# Reconnaissance automatique des entités nommées arabes et leur traduction vers le français

Hela Fehri

**UNIVERSITE DE FRANCHE-COMTE**
ECOLE DOCTORALE «LANGAGES, ESPACES, TEMPS, SOCIETES»

**UNIVERSITE DE SFAX (TUNISIE)**
ECOLE DOCTORALE DES SCIENCES ECONOMIQUES, GESTION ET INFORMATIQUE

**THESE** :
En vue de l'obtention du titre de docteur en :
**Sciences du langage (Université de Franche-Comté)**

En vue de l'obtention du titre de docteur en :
**Informatique (Université de Sfax, Tunisie)**

# RECONNAISSANCE AUTOMATIQUE DES ENTITES NOMMEES ARABES ET LEUR TRADUCTION VERS LE FRANÇAIS

Présentée et soutenue publiquement par

**Héla FEHRI**

Le 17 décembre 2012

Sous la direction de M. le Professeur Max SILBERZTEIN
M. le Professeur Abdelmajid BEN HAMADOU

Membres du jury :

Lamia BELGUITH, Maître de Conférences à l'université de Sfax, Membre
Abdelmajid BEN HAMADOU, Professeur à l'université de Sfax, Directeur
Kais HADDAR, Maître de conférences à l'université de Sfax, Membre
Laurent ROMARY, Professeur à l'université de Humboldt, Rapporteur
Max SILBERZTEIN, Professeur à l'université de Franche-Comté, Directeur
Yahia SLIMANI, Professeur à l'université de Tunis El Manar, Président du Jury
Rim ZITOUNI FAIZ, Professeur à l'université de Carthage, Rapporteur

# Table de matières

# Liste des figures

# Liste des tableaux

# Introduction générale

Avec l'évolution rapide des technologies de l'information et de la communication, le besoin s'est rapidement fait sentir de s'appuyer sur les techniques linguistiques pour faciliter la communication homme-machine. Parallèlement, la linguistique a pu profiter de la puissance des ordinateurs pour acquérir une nouvelle dimension et ouvrir la voie à de nouveaux domaines de recherche. Parmi ces domaines figure la traduction automatique. Cette dernière, longtemps sous-estimée, s'est en fait avérée l'une des tâches les plus délicates à effectuer par un ordinateur. Aux phases lexicales et syntaxiques, à peu près maîtrisées, s'ajoute une analyse sémantique, puis pragmatique, qui tente de déterminer le sens particulier d'un mot, dans le contexte où il apparaît. Le contexte lui-même pouvait s'étendre à l'ensemble du texte traduit.

Notons que, entre analyse lexicale et analyse syntaxique, il existe un niveau intermédiaire aux limites non précises formé d'un ensemble de phénomènes locaux (ex. : flexion de mots, fin de phrase, expressions figées, semi-figées).

En effet, selon (Gross, 1996), les expressions figées sont des : "Unités polylexicales présentant un caractère figé définies selon deux types de contraintes : syntaxique (liberté restreinte) et sémantique (opacité)". Certaines expressions dans la langue sont strictement figées (p. ex. : pomme de terre) : on peut en fait les classifier comme "mots composés". Par contre, d'autres ne sont que partiellement figées : elles n'acceptent pas n'importe quel complément (on voit clairement apparaître des contraintes sémantiques), mais elles offrent une certaine possibilité : on pourrait ainsi parler du directeur de la petite compagnie, du directeur de la thèse de doctorat,….On les appellera expressions semi-figées ou aussi **Entités Nommées (EN)**.

Une EN est donc une unité linguistique qui désigne un élément précis de l'univers du discours et qui peut-être un nom propre (Abdelmajid, Tunisie), un mot polylexical (le chef du département) mais également une mesure (un prix, par exemple) ou encore une date. Les EN désignent le plus souvent les éléments sur lesquels porte le discours. Leur détection est donc essentielle dans les applications d'extraction ou de recherche d'information textuelle.

Les EN sont particulièrement importantes pour l'accès au contenu du document car elles forment les briques élémentaires sur lesquelles repose l'analyse. Elles sont généralement considérées comme directement référentielles : ce sont les désignateurs rigides de Kripke

(1982) qui font référence aux objets du monde, organisés en ontologie. Ces séquences référentielles sont primordiales pour beaucoup d'applications linguistiques, que ce soit la recherche ou l'extraction d'information, la traduction automatique ou la compréhension de textes. Leur particularité est que leur reconnaissance nécessite la mise en œuvre des techniques portant sur des mots inconnus analysés en contexte (on ne dispose pas par exemple de dictionnaires complets de noms de personnes).

L'intérêt pour les EN est lié au développement des recherches sur corpus réels, notamment sur le web à partir du début des années 1990. Les conférences d'extraction d'informations (notamment les conférences américaines MUC « Message Understanding Conferences ») ont contribué à dynamiser et structurer ce domaine.

La traduction des EN d'une langue à une autre ouvre de nouvelles perspectives car elle peut être à la base de nouvelles applications notamment dans les domaines de l'accès multilingue aux informations, l'annotation/l'indexation des documents et l'enseignement à distance des langues. Cependant, la traduction ne conduit pas toujours aux résultats attendus surtout pour des langues de familles différentes. Pour cette raison, la translittération d'une partie d'une EN se révèle parfois nécessaire. La modélisation formelle ou semi-formelle des EN peut intervenir dans les deux processus de reconnaissance et de traduction. En fait, elle permet de rendre plus fiable la constitution des ressources linguistiques, de limiter l'impact des spécificités linguistiques et de faciliter les transformations d'une représentation à une autre pour la traduction. Toutefois, l'élaboration d'une représentation formelle n'est pas une tâche triviale car, d'une part, il faut trouver une représentation qui prend en compte la notion de récursivité et de longueur d'EN. D'autre part, elle doit contenir aussi un nombre suffisant de traits capables de décrire n'importe quelle EN indépendamment du domaine et de la catégorie grammaticale. Autrement dit, les mêmes traits doivent satisfaire tous les types d'EN.

Pour certaines EN, telles que les dates (Maurel, 1990), il paraît impossible de répertorier individuellement l'ensemble des constructions possibles. La combinatoire rend en effet le nombre de celles-ci trop important. Une représentation sous forme d'automates est bien plus efficace. Les modèles à états finis ont été appliqués, dans ce contexte, avec un certain succès (Maurel, 1990). Le formalisme des grammaires locales [ (Gross, 1993) et (Gross, 1997)] (voir Annexe 1), appelé aussi grammaires lexicalisées, est ensuite apparu comme une évolution très intéressante de ces modèles du fait de sa simplicité et sa modularité. Les grammaires locales sont équivalentes à des réseaux récursifs de transitions (Woods, 1970).

L'intérêt majeur des grammaires locales est de représenter de manière simple et compacte des contraintes lexico-syntaxiques définissant des classes syntaxiques comme les déterminants nominaux (Silberztein, 2003), les complexes verbaux (Gross, 1999) et même des classes syntaxico-sémantiques comme les adverbes de dates (Maurel, 1990), les prépositions locatives (Constant, 2003). À un moindre niveau, les grammaires locales sont aussi utilisées pour l'analyse de surface basée sur des contraintes grammaticales ou graphiques (i.e., reconnaissance d'EN (Friburger & Maurel, 2004)). Elles sont à la base de plusieurs plateformes linguistiques telles qu'Intex (Silberztein, 1993) et NooJ (Silberztein, 2005). Ces plateformes offrent un cadre de travail unifiant formats et formalismes.

Par ailleurs, en combinant de manière cohérente plusieurs grammaires locales entre elles, il devient possible d'effectuer une analyse syntaxique de grande précision sur de nombreux corpus. Ceci peut être utilisé, par exemple, pour la difficile tâche de désambigüisation : en élargissant le contexte du mot concerné aux unités lexicales qui l'entourent.

C'est précisément dans ce contexte que s'inscrivent nos travaux de recherche. Le principal objectif est de développer un système de reconnaissance et de traduction des EN arabes vers le français pour différents domaines. Le système à développer servira, d'une part, d'analyseur en vue d'une désambiguïsation sémantique des expressions retenues par le biais de grammaires locales, et d'autre part, d'un outil d'extraction et de traduction d'information du WEB. La plateforme avec laquelle le système de reconnaissance et de traduction est implémenté est celle de NooJ (Silberztein, 2005), une plateforme de TAL qui est à la fois un environnement de développement de ressources linguistiques et un outil pour le traitement automatique de corpus de grande taille. NooJ, comme son prédécesseur Intex, est basé sur la technologie à états finis : les textes, les dictionnaires et les grammaires y sont tous représentés par des transducteurs à états finis, d'où son efficacité. Il a été développé par Max Silberztein à l'Université de Franche-Comté depuis 1992. La théorie linguistique sous-jacente est celle de la grammaire transformationnelle de Z. Harris, adaptée à la description de la syntaxe française par M. Gross. Ainsi, la plateforme NooJ permet de valider et d'évaluer rapidement les idées et les démarches proposées.

Ainsi, notre première contribution dans ce domaine a consisté au raffinement effectué au niveau des catégories de la hiérarchie de type des EN proposée par des conférences MUC et à l'ajout d'autres catégories liées au domaine du sport [(Fehri et al., 2008) et (Fehri et al., 2010a)]. La deuxième contribution a consisté à proposer un modèle de représentation des EN [ (Fehri et al., 2010b) et (Fehri et al., 2011c)]. Ce modèle permet d'identifier les ressources

nécessaires, les niveaux de structuration des transducteurs et de favoriser la réutilisation. D'autre part, nous avons pu établir une démarche permettant d'utiliser les grammaires locales et plus précisément les transducteurs, dans les phases de reconnaissance et de traduction. Cette démarche est étendue par un module de translittération basé sur le système AL-Qalam (voir Annexe 3) et les spécificités phonologiques de la langue arabe [(Fehri et al., 2009a) (Fehri et al., 2009b)]. La troisième contribution s'est focalisée sur la séparation de la phase de reconnaissance de celle de la traduction afin d'assurer la réutilisation des ressources [(Fehri et al., 2011a) et (Fehri et al., 2011b)]. L'expérimentation du système développé a été effectuée en utilisant la plateforme linguistique NooJ. L'évaluation de la phase de reconnaissance a été réalisée à l'aide des métriques : rappel, précision et F-mesure. La phase de traduction a été évaluée en comparant les résultats obtenus par rapport à d'autres traducteurs connus : Babylon et Google.

Le travail réalisé et présenté dans ce rapport est structuré en six chapitres :

Dans le premier chapitre, nous discutons les différentes définitions attribuées à la notion d'EN. Nous présentons aussi quelques travaux existants sur la catégorisation des EN. En outre, nous décrivons les approches de reconnaissance des EN ainsi que les travaux basés sur chacune de ces approches. Nous terminons ce chapitre par la présentation des approches destinées pour la traduction des EN.

Le deuxième chapitre s'intéresse à la typologie des EN arabes et plus particulièrement celles appartenant au domaine du sport. En effet, nous commençons tout d'abord par présenter la définition que nous avons retenue pour la description d'une EN arabe valide. Ensuite, nous étudions les différentes formes d'EN arabes extraites du corpus d'étude. Puis, nous proposons un nouveau modèle typologique des EN arabes concernées. Enfin, nous terminons par la description de quelques phénomènes linguistiques rencontrés lors de l'étude du corpus.

Dans le troisième chapitre, nous proposons un modèle de représentation des EN arabes. Ce modèle s'inspire de la structure attribut/valeur et de la notion Tête/Expansion. Nous présentons aussi la structure et les traits proposés pour ce modèle ainsi que les principes de sa bonne formation. Enfin, nous montrons l'unification de deux représentations formelles et l'indépendance du modèle vis-à-vis de la langue et du domaine.

Le quatrième chapitre est consacré à la description de la démarche de reconnaissance des EN arabes et particulièrement les noms de lieux sportifs. Cette démarche est basée sur la hiérarchie des EN établie et sur le modèle de représentation des EN proposé. Nous commençons ce chapitre par une présentation générale de la démarche. Ensuite, nous

identifions les ressources nécessaires pour la reconnaissance des EN. Enfin, nous passons à la modélisation des ressources identifiées à travers l'utilisation des transducteurs.

Dans le cinquième chapitre, nous présentons la démarche proposée pour la traduction des EN arabes vers le français. Nous débutons ce chapitre par la présentation des problèmes liés à la traduction Arabe-Français des EN. Ensuite, nous donnons une présentation générale de la démarche proposée. Puis, nous détaillons les étapes du processus de traduction. Enfin, nous terminons par la description du processus de translittération destiné à l'amélioration du résultat de traduction.

Le sixième chapitre est consacré à la réalisation informatique tout en suivant les démarches décrites dans les deux chapitres précédents. Nous présentons dans un premier temps l'implémentation NooJ des ressources nécessaires pour la reconnaissance et la traduction des EN. Nous passons par la suite à la conception et à l'implémentation de l'outil de reconnaissance et de traduction. Pour ce faire nous avons choisi le langage UML pour la conception et le langage C# pour l'implémentation qui va permettre de valider les idées proposées. Enfin, nous expérimentons et nous évaluons l'outil développé sur un corpus de 4000 textes collectés de différents journaux arabes et sites web.

Ce travail est clôturé par une conclusion et des perspectives.

# Chapitre 1 : Etat de l'art sur la reconnaissance des EN et la traduction

Le présent chapitre est dédié à la présentation de quelques travaux en faveur de la reconnaissance des EN et de leur traduction automatique. Ces travaux traitent entre autres la catégorisation des EN en proposant de différentes classifications fondées sur la sémantique, la syntaxe ou la pragmatique. Les travaux liés à la manipulation des EN se basent sur une définition bien déterminée pour le concept d'EN. En effet, plusieurs définitions ont été adoptées. Ces définitions s'articulent autour de la couverture de l'EN et des critères qui doivent être vérifiés dans une entité pour qu'elle soit valide.

Les travaux sur la reconnaissance des EN utilisent l'une des approches suivantes : statistique, linguistique ou hybride. Ceux sur la traduction adoptent l'une des approches suivantes : experte ou empirique.

Nous débutons ce chapitre en présentant les différentes définitions attribuées à la notion d'EN. Ensuite, nous citons les catégorisations proposées pour la reconnaissance et la traduction automatiques des EN. Puis, nous décrivons les différentes approches de reconnaissance des EN ainsi que les travaux associés. Enfin, nous présentons les approches de traduction des EN.

## 1. Définitions de l'EN

Plusieurs définitions ont été attribuées à la notion d'EN. Dans ce qui suit, nous présentons quelques unes proposées dans le cadre des campagnes d'évaluation, des projets connus de reconnaissance d'EN et des « encyclopédies ». Le recensement de ces définitions permet de se faire une idée précise des constituants de base d'une EN dans les différents contextes.

Partant de ces définitions, nous constatons que les noms propres sont la base des EN. Mais, la différence entre-elles concerne le champ de couverture du nom propre. Autrement dit, est-ce que l'EN couvre uniquement les noms propres au sens classique ou aussi les noms collectifs, descriptifs etc. Il ne faut pas aussi oublier le différend concernant la question de l'unicité sémantique de l'EN mais encore la considération des expressions de temps et de quantité

comme des extensions pour la notion d'EN. Dans ce qui suit, nous citons les définitions existantes tout en s'appuyant sur les critères de référence et d'unicité.

## 1.1. Définitions basées sur le critère de référence

La plupart des définitions que nous avons trouvées dans la littérature se basent sur le critère de référence d'une EN. Dans ce qui suit, nous citons trois définitions différentes respectant ce critère.

La définition de (Chinchor, 1998) et (Poibeau, 2003) considère une EN comme un nom propre appartenant à un ensemble restreint. En effet, dans cette définition, un nom propre ne peut être qu'un nom d'une personne (ex., *Pierre*), de lieu (ex., *Paris*) ou d'organisation (ex., *ONU*). Les noms de maladies, les noms collectifs (*les Français*, *les Néandertaliens*, etc.) ou encore les noms de personnages mythiques ou fictifs (*Hercule*, *Colombo*, etc.) etc. ne sont pas considérés comme des EN mais plutôt une simple entité.

Contrairement à la première définition, celle de (Le Meur et al., 2004), n'effectue pas de restrictions sur les noms propres mais exige la présence d'un domaine pour parler d'une EN. Dans cette définition, une EN doit se référer à une entité du monde concret dans certains domaines spécifiques notamment humains, sociaux, politiques, économiques ou géographiques (typiquement un nom propre ou un acronyme) telle que FIFA et UNICEF.

La troisième est bien plus large. En effet, elle ajoute aux noms propres les expressions temporelles ou numériques et désigne parfois les éléments de base pour une tâche donnée (par exemple, les noms de gènes dans le cadre de l'étude des textes de biologie). Cette définition est adoptée par plusieurs chercheurs tels que (Friburger, 2002), (Tran, 2006), (Daille et al., 2000), (Sekine et al., 2002), (Fourour & Morin, 2003) et citée dans des encyclopédies libres telles que Atalapédie et Wikipédia.

## 1.2. Définition basée sur le critère de l'unicité

Autres travaux ((Weissenbacher, 2003), (Rangel Vicente, 2005), (Enjalbert, 2005)) ont ajouté au critère de référence le critère de l'unicité. Ils considèrent que le fait qu'une entité est référentielle n'est pas une condition suffisante pour parler d'une EN. Il faut aussi que cette entité soit unique c'est à dire mono-référentielle. Ainsi, un nom propre dans le sens classique peut ne pas être une EN. Par exemple, si nous prenons le cas du prénom « Jean », il peut désigner plusieurs personnes qui ont ce même prénom. Ainsi, ce prénom n'est pas considéré

comme une EN. Pour qu'il soit le cas, on lui ajoute la filiation complète en utilisant par exemple le mot "بن" *ben* ou "ولد" *wild* "fils de" selon les traditions du pays concerné. C'est pourquoi, on considère que l'EN comme **un syntagme** qui réfère à un unique objet d'une réalité supposée (Weissenbacher, 2003). Par conséquent, une EN devient un nom non ambigu et univoque qui doit suivre des patrons syntaxiques bien déterminés.

Les définitions présentées ci-dessus influencent directement les travaux de catégorisation des EN, objet du paragraphe suivant.

# 2. Travaux sur la catégorisation des EN

La catégorisation des EN qui consiste à identifier les types des EN en les affectant à des catégories est une étape préliminaire consacrée à tout traitement automatique. En effet, elle permet de réduire la complexité des différentes tâches qui peuvent être effectuées sur les EN telles que la reconnaissance et la traduction. Cependant, la catégorisation n'est pas une tâche triviale dans la mesure où il est non seulement nécessaire de déterminer les catégories à reconnaître mais aussi les différents constituants de chacune d'elles. C'est pourquoi, hormis les conférences MUC (Message Understanding Conference), les autres campagnes telles que ACE (Automatic Content Extraction) et ESTER fixent par la suite des sous-catégories de plus en plus nombreuses, afin d'expliciter quelque peu les choses. Ainsi et même en s'inscrivant explicitement dans le sillage de la définition d'une tâche, il semble difficile de se calquer sur une catégorisation parce que l'observation d'entités dans les textes conduit naturellement à l'évocation et donc à la détermination de nouvelles catégories. Dans ce contexte, Ehrmann dans son travail (Ehrmann, 2008) considère que le nombre de catégories ne semble pas être le plus important. L'essentiel pour lui, pour penser à une catégorisation utilisable dans le contexte de reconnaissance d'EN, est de prendre en compte le domaine applicatif, de considérer la catégorisation comme un véritable enjeu et d'adopter une démarche méthodologique pour sa réalisation. Donc, Il n'existe aucune catégorisation idéale ni de solution pour y parvenir ; le mieux semble être de suivre la proposition de S. Sekine, « We believe that there is no ultimate solution, so we seek rather empirical solution », et de multiplier les sources d'inspiration.

Dans ce qui suit, nous décrivons les différentes catégorisations existantes selon des critères tels que la sémantique, la syntaxe et la pragmatique.

## 2.1. Catégorisation référentielle (sémantique)

Huit catégorisations des EN ont été proposées en s'appuyant sur la référence. Dans ce qui suit, nous présentons le mouvement de changements et d'extensions effectué sur la catégorisation des EN en partant de celles existantes.

### 2.1.1. Catégorisation de Paik

Dans (Paik et al., 1996), les auteurs proposent une classification des entités, regroupant EN et entités temporelles, réalisée à partir d'une étude du *Wall Street Journal* qui comporte 30 catégories divisées en 9 classes : **Géographique** (villes, ports, aéroports, îles, comtés ou départements, provinces, pays, continents, régions, fleuves, autres noms géographiques), **Affiliation** (religions, nationalités) changée par le nom **Appartenance** en 1996, **Organisation (**entreprises, types d'entreprises, institutions, institutions gouvernementales, organisations) **Humain (**personnes, fonctions), **Document (**documents), **Equipement** (logiciels, matériels, machines), **Scientifique** (maladies, drogues, médicaments), **Temporelle (**dates et heures) et **Divers (**autres noms d'EN). Les 8 premières classes couvrent 89 % des entités présentes dans le corpus d'étude de Paik *et al.* (1996).

### 2.1.2. Catégorisation de Wolinski

Dans (Wolinski et al., 1995), les auteurs ont défini une autre catégorisation référentielle comprenant une cinquantaine de thèmes pour permettre le classement automatique des dépêches de l'Agence France Presse. Cette catégorisation n'est malheureusement pas détaillée dans leur article.

### 2.1.3. Catégorisation des conférences MUC

Les conférences MUC ont proposé trois catégories selon les informations à identifier : **ENAMEX** qui désigne des noms propres qui peuvent faire référence à des ***noms de personne***, ***de lieu*** ou ***d'organisation***, **TIMEX** qui désigne des expressions temporelles divisées en ***dates*** et ***heures*** et **NUMEX** qui désigne des ***expressions numériques*** qui font référence à des ***pourcentages*** ou à des ***valeurs monétaires***. La classe ENAMEX (Entity Name Extraction) qui regroupe les noms propres est communément appelée la classe des EN.

Il est à noter que les entités prises en compte par les systèmes de reconnaissance développés dans le cadre des conférences MUC ne considèrent pas toute la palette des entités intéressantes en TALN : les noms de médias, d'évènements, de documents, etc.

Bien que cette catégorisation regroupe une grande partie des EN présentes dans les textes journalistiques, elle est limitée et inadaptée à la traduction, car elle reste insuffisamment exhaustive et fine.

### 2.1.4. Catégorisation de la conférence IREX

*Irex* (Sekine & Isahara, 1999), organisée au Japon, reprend les 7 catégories énumérées dans les conférences MUC, avec cependant une catégorie supplémentaire nommée **Artifact**, pour annoter, entre autres, les noms de produits (Pentium Processor) ou de prix (Nobel Prize). De plus, une étiquette **optional** était prévue afin d'annoter une entité pour laquelle il serait difficile de déterminer la catégorie.

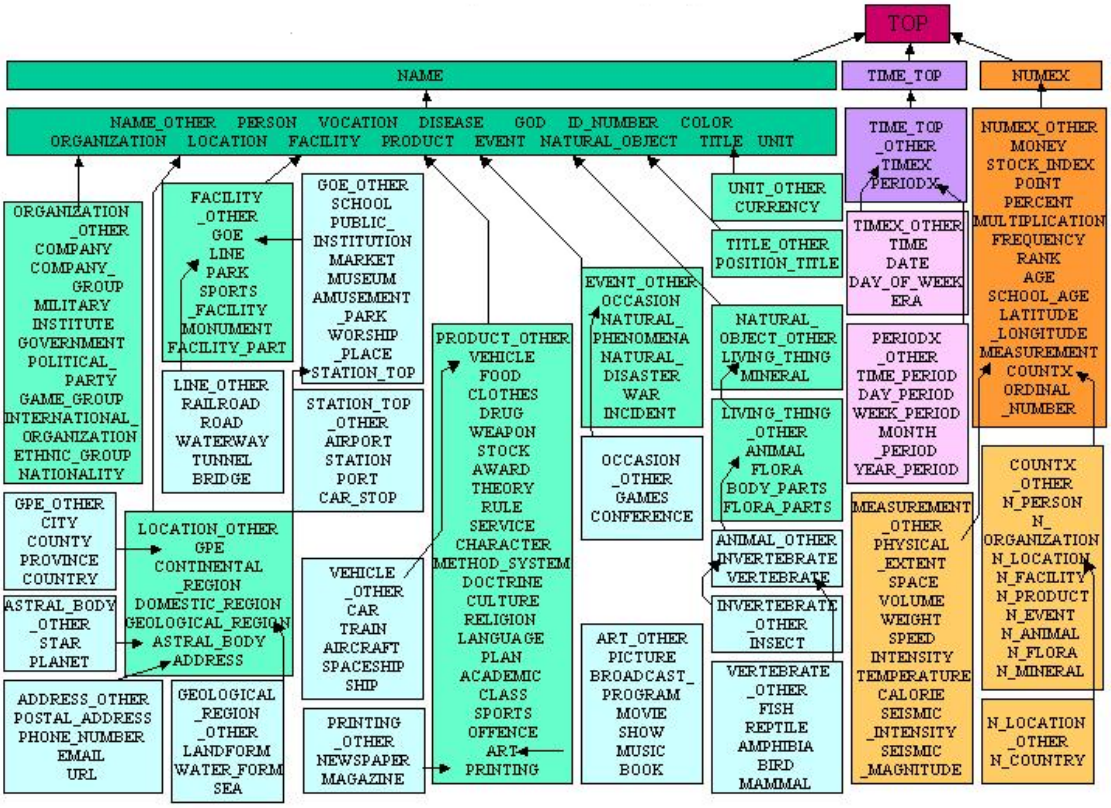


**Figure 1.** *Hiérarchie des EN de S.Sekine (version 6.1.2)*

Quelques années plus tard, une classification étendue est conçue afin de couvrir le plus d'EN possibles. Elle regroupe l'ensemble des EN définies par MUC (Grishman & Sundheim, 1996) et l'ensemble des EN développées par IREX (Sekine & Eriguchi, 2000). Au départ, 150

catégories ont été définies (Sekine et al., 2002), ensuite 50 catégories ont été ajoutées (Sekine & Nobata, 2004) (Sekine, 2004). Ces travaux sont pour des applications générales (ex., système question(s)-réponse(s), extraction d'information). La Figure 1 présente la hiérarchie complète de S.Sekine.

### 2.1.5. Catégorisation de la conférence CoNLL

Les conférences CoNLL 2002 et 2003 (Tjong Kim Sang, 2002) et (Tjong Kim Sang & De Meulder, 2003) procèdent aussi à quelques changements par rapport à la catégorisation de MUC. En effet, elles reprennent uniquement les catégories **Personne**, **Lieu** et **Organisation** et ajoutent la catégorie **Miscellaneous** pour annoter les entités n'appartenant pas aux classes de MUC.

### 2.1.6. Catégorisation de la compagne ACE

La première modification importante de la catégorisation de MUC fut apportée durant la campagne ACE (ACE Pilot Study Task Definition, 2000) et (ACE, 2005). En effet, dans ce projet, il ne s'agit pas de reconnaître des EN mais des entités, c'est-à-dire l'ensemble des réalisations lexicales des EN, soit des noms propres, des noms communs et des pronoms. La tâche pilote de ACE prévoyait les catégories suivantes : **Personne**, **Organisation**, **Bâtiments** (facility), **Lieux** et **GPE**. Cette dernière signifie « Geo-Political Entities », et concerne les entités géographiques également définies par des aspects politiques et sociaux. La catégorie GPE, non prévue au départ, est ajoutée afin de distinguer entre une entité strictement géographique (ex., Tunis, Paris) et une autre ayant un aspect politique et social (ex., Union Européenne). Et comme, une entité peut référer simultanément à plusieurs catégories, la catégorie GPE a été créée pour rendre compte de la sémantique de certaines entités et éviter bien des hésitations lors de l'annotation. En définitive, lors de la définition de la tâche réelle, deux catégories supplémentaires furent encore ajoutées : **Véhicule** *et* **Arme**. Peu de temps après, la compagne ACE organise conjointement avec plusieurs universités scandinaves le projet Nomen Nescio. L'objectif est de reconnaître, classifier et annoter des noms dans des textes courants. Dans ce projet, seules les catégories « universelles » sont présentes et trois autres sont ajoutées : **Event**, O**bject** et **Work&Art**.

### 2.1.7. Catégorisation de la compagne ESTER

La campagne française ESTER a imposé la catégorisation des EN en huit catégories, qui sont: **Personne**, **Organisation**, **Groupe geo-Socio-Politique** (GSP), **Lieu**, **Bâtiment**, **Produit**, **Temps** et **Quantité**, auxquelles est ajoutée, en cas d'incertitude, la catégorie **Inconnu** ; l'héritage de ACE se fait ici sentir au travers des catégories **GSP** (proche de GPE) et **Bâtiment**, celui de IREX au travers de la catégorie **Produit**.

### 2.1.8. Catégorisation de la compagne HAREM

Quant à la campagne portugaise HAREM, elle a proposé, elle aussi, un nombre relativement conséquent de catégories. Aux trois fondamentales sont ajoutées les catégories **Evénement**, **Chose**, **Œuvre**, **Abstraction**, **Temps**, **Valeur**, ainsi que le désormais habituel **divers**. Cette multiplication des catégories est due avant tout à la confrontation des annotateurs avec des textes et traduit une volonté de généraliser.

## 2.2. Classification graphique (syntaxique)

La distinction des EN, suivant des critères graphiques, est intéressante dans une optique de reconnaissance automatique. Suivant la graphie, l'identification et la classification des EN entraîneront des traitements différents. Nous distinguons, ainsi, les catégories suivantes proposées par (Daille et al., 2000) et inspirées de la terminologie de (Jonasson, 1994) :

- **EN pures simples :** se sont les EN constituées d'une seule unité lexicale commençant par une majuscule comme France, Aristote ;
- **EN pures complexes :** se sont les EN constituées de plusieurs unités lexicales commençant par une majuscule comme Conflans Saint-Honorine. Aussi, une sous-catégorie **Prénom Nom** est proposée. Cette sous-catégorie couvre les EN constituées d'un prénom (ou plusieurs prénoms) et d'une unité lexicale commençant par une majuscule référant à un nom de personne comme Paul Valéry ;
- **EN faiblement mixtes :** se sont les EN constituées de plusieurs mots commençant par une majuscule et contenant des mots de liaison en minuscule comme le Jardin des Plantes. Cette liste de mots de liaison est fermée et comprend des articles, des prépositions et des conjonctions de coordination, etc. ;

- **EN mixtes :** se sont les EN constituées de plusieurs unités lexicales dont au moins une commence par une majuscule comme le Comité international de la Croix-Rouge, le Mouvement contre le racisme et pour l'amitié entre les peuples ;
- **Sigles** : se sont les EN constituées d'une seule unité lexicale comportant plus d'une majuscule et dont chaque lettre en majuscule réfère elle-même à une autre unité lexicale comme USA. Il est à noter qu'il est important de distinguer, au niveau graphique, les EN (appartenant à cette catégorie) qui réfèrent à des EN pures complexes de celles qui réfèrent à des EN mixtes (faibles ou non).

## 2.3. Catégorisation basée sur des aspects pragmatiques

La seule classification basée sur des aspects pragmatiques et destinée pour la traduction, d'après (Daille et al., 2000), est celle réalisée par le linguiste germanophone Bauer (Bauer, 1985). Ce linguiste énumère ce qui, par convention, constitue un nom propre en prenant en considération des éléments extralinguistiques propres au référent. Sa typologie est constituée de six classes principales, avec pour chacune, plusieurs catégories :

- **anthroponymes** : cette classe représente les personnes individuelles ou les groupes tels que les patronymes, les prénoms, les *pseudonymes,* les gentilés, les hypocoristes, les *ethnonymes,* les groupes musicaux modernes, les ensembles artistiques et orchestres classiques, les partis et les organisations.
- **Toponymes** : cette classe recouvre les noms de lieux tels que les pays, les villes, les microtoponymes, les hydronymes, les oronymes et les installations militaires.
- **Ergonymes** : cette classe représente les objets et les produits manufacturés et par extension les marques, les entreprises, les établissements d'enseignement et de recherche, les titres de livres, de films, de publications et d'œuvres d'art.
- **Praxonymes** : cette classe englobe les faits historiques, les maladies et les événements culturels.
- **Phénonymes** : cette classe recouvre les ouragans, les zones de haute et de basse pressions, les astres et les comètes.
- **Zoonymes** : cette classe désigne les noms d'animaux familiers.

Grass (Grass, 2000) a suivi la même typologie que celle du Bauer (1985) ; il a éliminé juste la classe Zoonymes.

Hormis la classe des entités temporelles, il existe de nombreuses similitudes entre la catégorisation de Paik et al. (1994) et celle de Bauer (1985). Néanmoins, certaines classes de Bauer (1985) comme les praxonymes ou les phénonymes n'apparaissent ni dans les classes, ni dans les catégories de Paik et al. (1994, 1996). Inversement, toutes les catégories présentes dans Paik et al. (1994, 1996) peuvent s'insérer dans les classes de (Bauer, 1985).

# 3. Approches de reconnaissance des EN

Les approches de reconnaissance d'EN ont connu un fort développement depuis la fin des années 80 sous l'impulsion des conférences MUC. Trois grandes approches sont généralement suivies : l'approche linguistique ou symbolique, l'approche statistique ou à base d'apprentissage et l'approche hybride. Ce qui distingue les approches citées, n'est pas la nature des informations prises en compte, mais plutôt leur acquisition et leur manipulation. Dans ce qui suit, nous donnons un aperçu de ces approches tout en s'appuyant sur quelques exemples représentatifs des systèmes de reconnaissance d'EN.

## 3.1. Approche linguistique

L'approche linguistique repose sur l'intuition humaine, avec la construction manuelle des modèles d'analyse, le plus souvent sous la forme de règles contextuelles. C'est pourquoi, cette approche est appelée aussi approche à base de règles. Ces dernières, qui expriment l'information à reconnaitre, prennent la forme de patrons d'extraction permettant la description d'enchaînements possibles de syntagmes nominaux. Ces patrons exploitent généralement des informations d'ordre morphosyntaxique telles que les mots déclencheurs (السيّد *Mr*, ملعب *stade*), ainsi que celles contenues dans des ressources (lexiques ou dictionnaires).

Ce type d'approche fut largement répandu pour la reconnaissance des EN, voire majoritaire durant les années 1990, au temps des premières conférences MUC avant que l'apprentissage ne fasse son apparition dans le domaine (MUC-6, 1995). Un système de reconnaissance d'EN basé sur l'approche linguistique comporte par exemple les règles suivantes : si un mot déclencheur indiquant un nom de lieu précède un nom d'une personnalité (connaissance issue d'un dictionnaire), alors le syntagme peut être étiqueté comme un nom de lieu ; ou bien : si le mot فريق *faryq* (Equipe) est suivi d'une nationalité (connaissance issue d'un dictionnaire), alors il s'agit d'un nom d'une équipe. En fait, nous recourons à l'utilisation de ces preuves

(ex., mot déclencheur) pour la détection du début d'un nom propre en langue arabe comme le cas de présence de majuscule au début des noms en français et les autres langues romanes. Notons que, la présence de ces preuves facilite la reconnaissance mais ne résout pas tous les problèmes notamment celui de l'identification de la frontière droite de l'EN.

Dans ce cadre, (Mesfar, 2008) a défini l'architecture d'un système de reconnaissance des EN arabes. Ce système est fondé sur une approche linguistique utilisant les grammaires locales implémentées dans la plateforme linguistique NooJ. L'architecture de ce système (Mesfar, 2008) est illustrée dans la **Figure 2**.

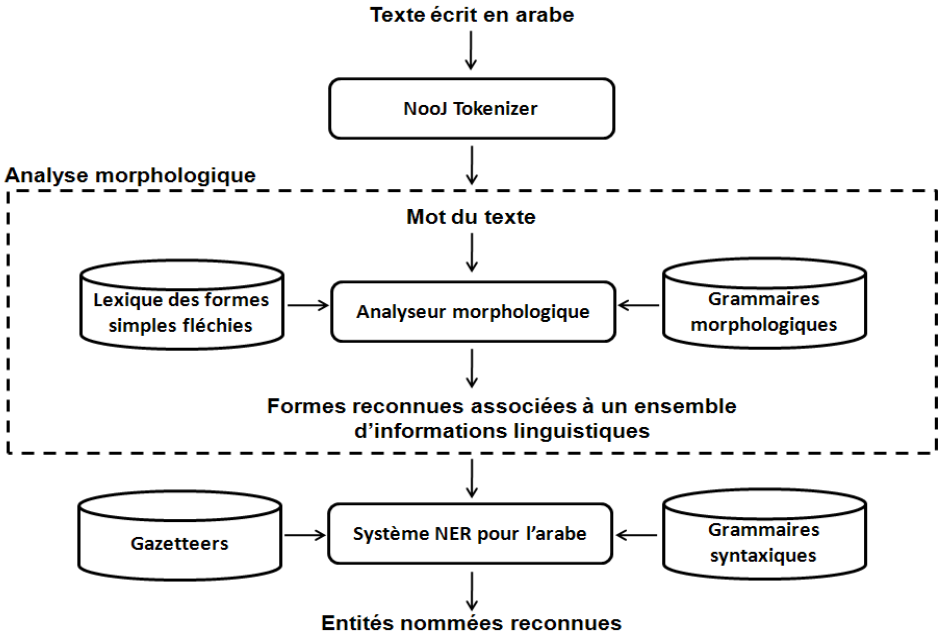


**Figure 2.** *Architecture générale d'un système de reconnaissance des EN*

Comme indiqué dans la **Figure 2**, le système de (Mesfar, 2008) passe par une analyse morphologique suivie d'une analyse syntaxique. L'analyse morphologique, qui se sert des dictionnaires et des grammaires morphologiques, permet l'identification des caractéristiques de chaque mot du texte existant dans les dictionnaires construits et la résolution de certains problèmes liés à la langue arabe comme l'agglutination. L'analyse syntaxique permet la reconnaissance des séquences pertinentes via l'application des grammaires syntaxiques et la production en résultat de certaines informations linguistiques, notamment le type de l'EN identifiée (nom de personne, organisation, localisation, etc.). Ces séquences sont décrites par le biais de règles manuellement construites.

Une évaluation de ce système est effectuée sur un corpus extrait du journal « Le Monde Diplomatique » donne, d'après l'auteur, des résultats satisfaisants. Les problèmes rencontrés

dans ce travail concernent essentiellement l'absence de conventions pour l'écriture des noms propres et la délimitation des entités identifiées.

Dans le même contexte, (Shaalan & Raza, 2009) ont développé un système de reconnaissance des EN arabes (NERA) en utilisant une approche fondée sur des règles. Les ressources créées sont : une liste blanche représentant un dictionnaire des noms et une grammaire, sous la forme d'expressions régulières, qui sont responsables de reconnaître les EN. Un mécanisme de filtrage est utilisé. Ce mécanisme sert à deux fins différentes :

a- la révision des résultats d'un extracteur d'EN à l'aide de métadonnées, en termes d'une liste noire ou « rejecter », des EN mal formées.

b- La désambigüisation des sélections identiques ou présentant un chevauchement textuel trouvées par les différents extracteurs de l'EN pour obtenir le bon choix.

Dans le système NERA, les auteurs ont abordé aussi les défis majeurs posés par la reconnaissance des EN dans la langue arabe. NERA a été effectivement évalué en utilisant des corpus étiquetés.

Pour la langue française, nous pouvons citer le système d'extraction linguistique des noms propres en français baptisé ExtractNP (Friburger, 2002). Ce système est fondé sur les cascades des transducteurs. Grâce à celles-ci de multiples transformations sont réalisés sur les textes soit au niveau de l'analyse syntaxique, soit au niveau de l'extraction de l'information. ExtractNP utilise le système CasSys pour la génération de deux cascades de transducteurs d'extraction des noms propres : la première cascade a pour objectif de catégoriser et extraire une partie des noms propres d'un texte c'est-à-dire ceux qui possèdent des indices permettant de les repérer. La seconde cascade exploite les résultats de la première cascade afin d'en trouver de nouveau. Cette dernière apporte une bonne amélioration pour les résultats d'extraction.

Dans le système ExtractNP de nombreuses ressources linguistiques ont été utilisées pour l'extraction. Pour le repérage des organisations, l'auteur a utilisé des preuves externes et internes. Pour l'extraction des noms des lieux, elle tend à extraire les preuves internes d'un côté et d'utiliser des dictionnaires sans preuves de l'autre côté. Pour la localisation des noms de personne, elle a opté pour l'utilisation des preuves internes et externes outre l'exploit de la morphologie des prénoms et des patronymes. Le système ExtractNP fournit actuellement de bons résultats sur la langue française. En effet, il obtient un rappel de 93% et une précision de 94%.

En outre, et toujours pour la langue française, nous mentionnons le projet EPIDEMIA (Roux et al., 2006). Ce dernier se base sur l'utilisation des dépêches pour analyser la situation épidémiologique. La démarche d'extraction dans ce projet peut être résumée en deux étapes : la première consiste en l'analyse locale ayant pour objectif la reconnaissance des EN telles que les noms des hôpitaux et des pathologies. Cette analyse est réalisée par l'utilisation des transducteurs à états finis dans NooJ. Dans la deuxième étape, les partenaires du projet EPIDEMIA ont eu recours à la normalisation dont l'objectif est l'enrichissement des structures.

Nous terminons par d'autres travaux qui se focalisent sur l'utilisation de la morphologie pour le repérage des noms propres. A titre d'exemple, nous signalons le projet d'extraction des noms propres en anglais réalisé par (Coates-Stephens, 1992). Ce dernier utilise un certain nombre d'heuristiques morphologiques. Comme exemple de règles utilisées : *si un mot se termine par « ese », « ians » ou « «ian » alors c'est un nom d'origine. Par contre si un mot se termine par «shire » alors c'est un nom de lieu*. Cette règle est utilisée dans le cas où la racine d'un mot est un nom de lieu. D'autres travaux utilisent la morphologie pour l'extraction des noms de compagnies sous forme de sigles dans les textes (Taghva & Gilbreth, 1999). En effet, un candidat de sigle est composé de 3 à 10 lettres tel que l'acronyme d'organisation RTL.

## 3.2. Approche statistique

L'approche statistique a pour principe de base la mise au point automatique de modèles d'analyse à partir de volumes importants de données (ou corpus). Ces méthodes sont dites statistiques ou à base d'apprentissage car elles apprennent, à partir de corpus annotés, des modèles d'analyse de textes. Ces derniers peuvent prendre différentes formes comme arbres de décision, ensembles de règles logiques, modèles probabilistes ou encore chaînes de Markov cachées. Au regard de la reconnaissance d'EN, un système *« observant »* plusieurs fois la présence de l'abréviation « Mme » devant un mot annoté comme nom de personne (dans le corpus d'apprentissage) pourra facilement en déduire un modèle d'analyse pour ce type. Ces systèmes à base d'apprentissage se sont considérablement multipliés ces dernières années, eu égard à leur facilité de mise en œuvre.

Nous commençons tout d'abord avec le système de reconnaissance d'EN arabes ANERsys développé par (Benajiba, 2009). Ce système possède deux versions différentes basées sur une approche d'entropie maximale. La première version passe par une seule étape comportant deux phases. Ces deux phases sont illustrées dans la **Figure 3**.

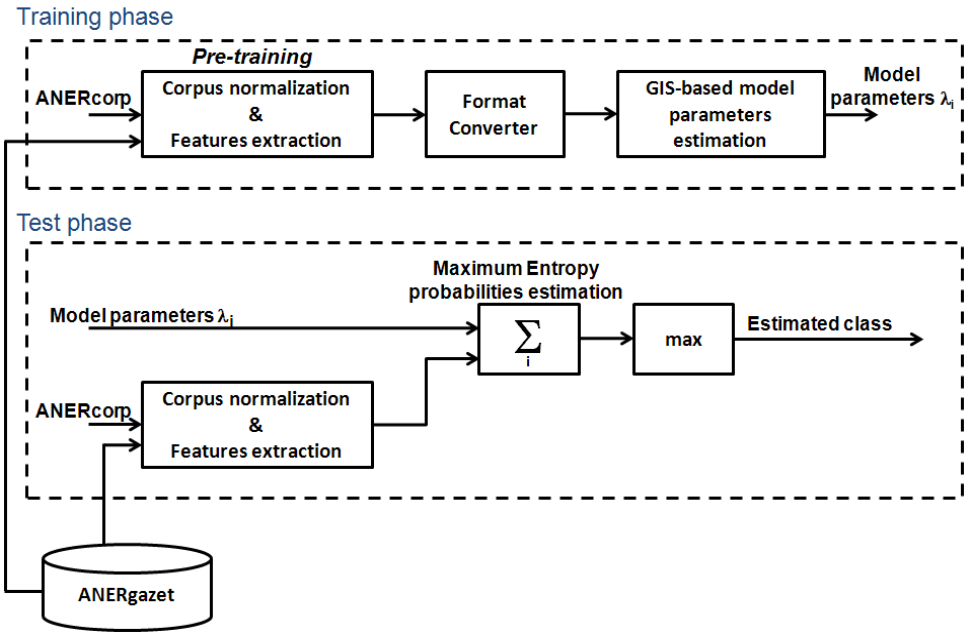


**Figure 3.** *Architecture de la première version de ANERsys*

La première phase est composée de trois modules.

**Le module de pré-apprentissage :** Il permet la normalisation des caractères. Chaque caractère « 'alef » avec « hamza » sera remplacé uniquement par « 'alef». Ce module est important vu que dans des documents différents nous pouvons trouver le même mot écrit de deux façons différentes (ex., أمير *'amyr* (prince) et امير *aamyr* (prince)). Ensuite, une préparation de traits (mot courant, contexte, etc.) est lancée afin de calculer toutes les fréquences nécessaires uni/bi-grammes et aussi de vérifier les mots existant dans ANERgazet, etc.

**Le module d'estimation des paramètres** : Il permet le calcul du poids des traits $\lambda_i$ en utilisant l'algorithme GIS (Generalized Iterative Scaling) et en utilisant l'outil YASMET « Yet Another Small MaxEnt Toolkit ». Ce dernier est développé en C++ et est téléchargeable gratuitement.

**Le module de conversion de format :** Ce dernier est inclut pour effectuer une conversion afin de présenter les traits, qui ont été préparés dans le premier module, dans le format requis par YASMET.

La deuxième phase est celle de test. Elle a pour objectif l'utilisation des paramètres obtenus dans la première phase afin de calculer la probabilité de chaque mot appartenant à chaque classe. Chaque nom de ces classes est précédé par B pour désigner le premier

mot (début) d'une EN ou par I pou désigner les autres mots d'une EN. Les mots qui ne font pas partie de l'EN sont représentés par O. Cette phase de test nécessite trois modules.

**Le module de pré-apprentissage** afin de collecter les traits pour chaque mot $w_i$.

**Le module d'estimation des paramètres** en utilisant l'entropie maximale. Ce module permet de calculer la probabilité de chaque mot $w_i$ dans chaque classe.

**Le module max** est un petit script qui permet de rechercher la classe ayant la probabilité maximale et d'attribuer cette probabilité au mot $w_i$ concerné.

L'évaluation de la première version de ANERsys est effectuée sur un corpus collecté de différents journaux, magazines et sites web tels que http://www.aljazeera.net ,http://www.raya.com et http://ar.wikipedia.org. Les résultats obtenus sont illustrés dans le Tableau **1**.

| | **Précision** | **Rappel** | **F-mesure** |
|---|---|---|---|
| **Location** | 82.17% | 78.42% | 80.25 |
| **Miscellaneous** | 61.54% | 32.65% | 42.67 |
| **Organization** | 45.16% | 31.04% | 36.79 |
| **Person** | 54.21% | 41.01% | 46.69 |
| **Overall** | 63.21% | 49.04% | 55.23 |

**Tableau 1.** *Evaluation du système ANERsys version 1*

Ces résultats montrent que l'approche de l'entropie maximale et l'ensemble de traits utilisés contribuent significativement pour capturer les EN qui n'ont pas été vues dans la phase d'apprentissage. Toutefois, le problème majeur réside dans le cas où une EN est composée de plusieurs mots.

Par exemple, dans le cas de l'EN « الرئيس السوري بشار بن حافظ الأسد » *elraaeys elsury bashshaar bin Haafidh al'asad (Le président syrien Bcahar fils de Hafedh alassad)*, ANERsys annote cette EN de la façon suivante : الرئيس *elraaeys (président)* O, السوري *elsury (syrien)* O, بشار *bashshaar (bachar)* B-PERS, بن *bin* (*fils de)* B-PERS, حافظ *Haafidh* (*Hafedh*) O, الأسد *al'asad (alassad)*, O. Cela montre que ANERsys ne peut reconnaître que deux mots : بشار *bashshaar (Bachar)* et بن *bin* (*fils de)* ; le premier mot بشار est reconnu vu qu'il vient juste après une nationalité (trait existant) et le deuxième apparaît fréquemment comme une partie d'une EN de type personne donc normalement, il doit être annoté comme I-PERS mais vu que le mot qui le précède est mal classifié, ANERsys l'annote comme B-PERS.

Pour résoudre ce problème, Benajiba a divisé la tâche de reconnaissance en deux sous-tâches. La première permet uniquement la détection des EN dans un texte et la deuxième exploite les

traits générés, par la première, pour classifier les EN. La **Figure 4** illustre ces deux sous-tâches.

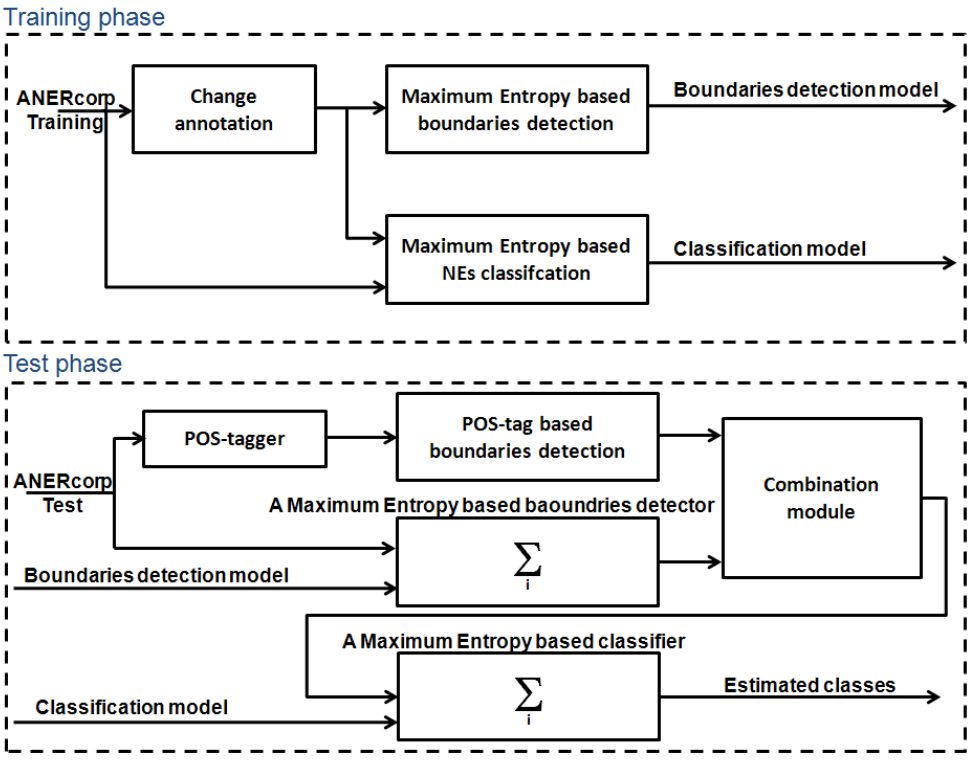


**Figure 4.** *Architecture générique de la deuxième version de ANERsys*

La phase d'apprentissage décrite dans la **Figure 4** comprend deux modules. Le premier module permet de détecter les limites d'une EN. La réalisation de ce module nécessite le changement des annotations B-PERS, B-LOC, B-ORG et B-MISC par B-NE et les annotations I-PERS, I-LOC, I-ORG et I-MISC par I-NE. Ainsi, le modèle obtenu est apprenti sur uniquement deux classes : I-NE et B-NE. Les données, les traits et l'outil d'apprentissage sont les mêmes que ceux utilisés dans la première version de ANERsys.

Le deuxième module de la phase d'apprentissage permet la classification des EN. Ce module utilise I-NE et B-NE comme des traits.

La phase de test comprend plusieurs modules ayant des comportements plus compliqués. Cette phase permet, dans une première étape, de détecter les limites des EN en utilisant un processus qui consiste à associer aux mots d'un texte leur fonction grammaticale (ce processus est téléchargeable gratuitement (POS-tagger)) et en calculant l'entropie maximale qui utilise les limites de l'EN détectées au niveau de la phase d'apprentissage. Dans une deuxième étape, la phase de test permet de classifier les EN déjà délimitées. Les résultats obtenus par la deuxième version de ANERsys sont illustrés dans le **Tableau 2**.

| | Précision | Rappel | F-mesure |
|---|---|---|---|
| **Location** | 91.69% | 82.23% | 86.71 |
| **Miscellaneous** | 72.34% | 55.74% | 62.96 |
| **Organization** | 47.95% | 45.02% | 46.43 |
| **Person** | 56.27% | 48.56% | 52.13 |
| **Overall** | 70.24% | 62.08% | 65.91 |

**Tableau 2.** *Evaluation du système ANERsys version 2*

Il est à noter que ANERcorp a été examiné à plusieurs reprises pour assurer la cohérence de ses annotations. De plus, Benajiba a réalisé des modèles basés sur SVM (Support Vector Machines) et CRF (Conditional Random Fields) qui sont prêts à être utilisés gratuitement et qui sont disponibles pour la communauté des chercheurs (http://www.dsic.upv.es/grupos/nle/downloads.html). Ces modèles peuvent être utilisés afin d'avoir un système de reconnaissance des EN arabes qui peut être ajusté à l'aide de l'étude fournie dans (Benajiba, 2009).

Également, parmi les systèmes basés sur l'approche statistique nous pouvons citer celui de (Frantzi, 1998). Ce système consiste à retenir comme termes candidats ceux qui possèdent la plus grande C-value représentant une valeur basée sur la fréquence d'apparition du mot et sur sa longueur. Alors que dans le travail de (Maynard & Ananiadou, 1999), cette valeur est combinée avec un facteur contextuel qui correspond à la prise en compte de la fréquence d'apparition des noms, des adjectifs et des verbes apparaissant dans le contexte du terme candidat. Pour le calcul de la « valeur terminologique » finale du terme candidat à partir d'un réseau sémantique, il faut calculer une distance entre les termes du contexte et le terme candidat. Les expériences décrites dans (Maynard & Ananiadou, 1999) s'intéressent au domaine médical et emploient le réseau sémantique UMLS (Unified Medical Language System) (NLM, 1997).

## 3.3. Approche hybride

L'approche hybride consiste à combiner l'approche linguistique et l'approche statistique afin de profiter des avantages des deux. En extraction d'informations, cette idée se traduit par l'association de la puissance descriptive des solutions linguistiques et des facilités statistiques d'apprentissage. Généralement, la partie essentielle de la méthode d'extraction est statistique. Quant à la partie linguistique, elle consiste à filtrer les termes en fonction de leur catégorie

syntaxique (Meilland & Bellot, 2005). Bien évidemment, l'utilisation d'information sémantique est envisagée bien qu'elle soit difficile à mettre en œuvre.

Parmi les projets admettant cette approche hybride, nous citons le système d'extraction des EN arabes proposé par (Elkateb-Gara, 2004). Sa méthodologie mixte passe nécessairement par trois étapes linguistiques et une étape statistique :

- L'analyse lexicale du texte ou éventuellement d'un corpus.
- La reconnaissance des séquences pertinentes à travers des grammaires dédiées qui traduisent les différentes règles de reconnaissance.
- L'étiquetage des séquences isolées afin d'avoir un texte annoté.

Les étapes linguistiques sont illustrées dans la **Figure 5**.

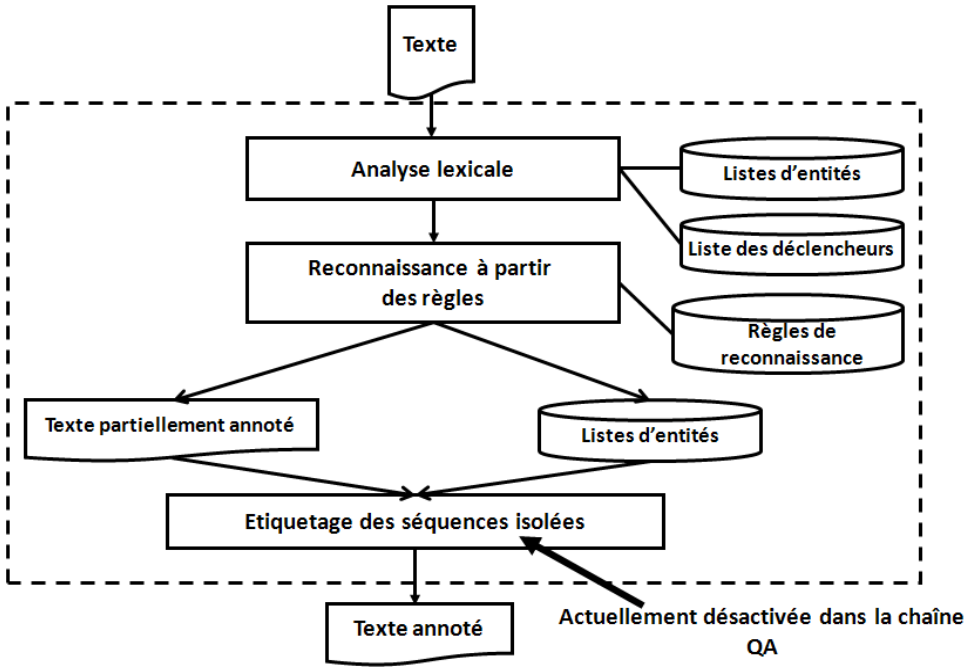


**Figure 5.** *Démarche proposée*

La première expérimentation de ce système (basé sur des traitements linguistiques) a relevé de multiples problèmes. Citons par exemple les problèmes de non traitement de séquences isolées, la mauvaise délimitation des EN reconnues et l'incomplétude que ce soit au niveau de la liste des déclencheurs, au niveau de la liste des EN ou au niveau des règles. Pour la résolution de ces différentes insuffisances, Faiza Elkateb a eu recours à une solution d'apprentissage pour étiqueter les séquences isolées. Elle a aussi, opté pour l'augmentation de nombre de règles ainsi que le nombre de déclencheurs afin d'avoir le système d'extraction le plus performant possible.

Egalement, dans ce cadre de méthodologie mixte, nous citons le système Nemesis (Fourour, 2002) qui consiste à réaliser une reconnaissance de noms propres à l'aide de lexiques de preuves externes et internes, associées à des étiquettes sémantiques, et encore à l'aide des expressions régulières, appliquées aux textes. Ce système est basé sur une catégorisation référentielle et une catégorisation graphique. Nemesis se fonde sur des règles de grammaire, il exploite des lexiques spécialisés et comporte un module d'apprentissage. Il se compose de quatre modules qui effectuent un traitement séquentiel immédiat des données : prétraitement lexical, première reconnaissance des EN, apprentissage et seconde reconnaissance des EN. Afin de compléter la reconnaissance, de nouveaux lexiques ont été créés automatiquement. Dans ce projet, Fourour a présenté des conflits de chevauchement, d'inclusion et d'accolement (c'est le cas où deux EN sont identifiées, l'une se situe immédiatement à la suite de l'autre) qui sont résolus par la suite soit par l'ajout de nouvelles règles, soit par le changement des priorités des règles ou par la fusion de deux règles.

Pour améliorer Nemesis, l'étude des conflits devra amener à mettre en place un véritable module de désambiguïsation. Parallèlement, elle devra permettre d'inférer de nouvelles règles durant la première reconnaissance et de concevoir un algorithme pour les insérer parmi les règles déjà existantes.

Il est à noter que dans Nemesis, Fourour a mis l'accent sur l'apport du web dans la reconnaissance des EN. En effet, il permet la catégorisation référentielle des EN non identifiées durant l'approche suivie.

Les performances atteintes par Nemesis, sur les anthroponymes et les toponymes, sont de 90% pour le rappel et 95% pour la précision.

Comme autre projet utilisant l'approche hybride, nous pouvons citer celui de (Frantzi et al, 2002) pour le Polonais, qui adopte l'approche C/NC-value. En premier lieu, Frantzi utilise des patrons syntaxiques pour identifier des termes candidats dans les textes annotés. En second lieu, ces termes candidats retenus sont catégorisés selon leur stabilité par rapport aux termes les plus longs (C-value) et aussi selon leur importance par rapport au contexte dans lequel ils apparaissent. L'évaluation de ce système sur les corpus d'entrainement et de test a relevé un taux de rappel 60,78% et de précision 66%. La performance du système est raisonnable mais elle pourrait être améliorée.

## 3.4. Discussion

Les avantages et les inconvénients respectifs des deux approches linguistique et statistique peuvent être résumés en ces points :

- l'approche linguistique reproche à l'approche statistique, entre autres, la non disponibilité de corpus annotés,

- l'approche statistique critique la première sur le temps nécessaire de développement ainsi que leur coût. Il est vrai qu'un travail de plusieurs mois d'un linguiste-informaticien est nécessaire pour l'écriture des règles, mais l'inverse est vrai également pour l'annotation de corpus qui peut être aussi long, même si cela peut se faire par des gens moins experts.

- La précision est plus importante pour les systèmes symboliques tandis que les systèmes à base d'apprentissage présentent l'avantage d'être plus flexibles mais aussi d'être plus robustes sur des corpus difficiles (ou bruités).

Enfin, au-delà de ces deux types d'approches et des désaccords de leurs partisans respectifs, il existe une troisième voie qui consiste à combiner l'approche symbolique et l'approche statistique en une approche qualifiée de mixte ou d'hybride. Cette dernière, rendue possible grâce à la maturité acquise par les deux autres, est sans doute la plus prometteuse.

# 4. Approches de traduction des EN

La traduction automatique des EN est une tâche pour laquelle la reconnaissance des EN constitue également une amorce importante. Il existe globalement deux grandes approches de base de TA : l'approche experte et l'approche empirique (Lavecchia, 2010). La première est fondée sur les connaissances d'experts humains. Elle renferme trois méthodes dérivées : la TA directe, la TA à base de règles de transfert et la TA fondée sur une interlangue. La deuxième est fondée sur l'extraction des connaissances à partir des quantités importantes de données textuelles. Elle peut être subdivisée en deux grandes familles : l'approche par analogie (ou à base d'exemples) et l'approche statistique. Ces deux dernières, contrairement aux méthodes de l'approche experte, ne nécessitent aucune connaissance a priori pour développer un système de traduction.

La TA des EN est effectuée dans la majorité des travaux en suivant l'approche statistique. Parmi ces travaux, nous citons par exemple celui de (Al-Onaizan & Knight, 2002). Ce travail consiste à traduire les EN arabes vers l'anglais en utilisant un algorithme basé sur des

ressources monolingues et bilingues. En effet, étant donné une EN dans la langue source, l'algorithme de traduction génère d'abord une liste de classement des candidats de traduction en utilisant les ressources bilingues et monolingues. Ensuite, la liste des candidats est recalculée à l'aide de différents indices monolingues. Dans le même contexte, nous trouvons le travail de (Ling et al., 2011) qui permet de récupérer une liste de documents web dans la langue cible, d'extraire les « anchor » textes à partir des liens de ces documents et de recenser la bonne traduction de l'EN à partir de ces textes en utilisant une combinaison de traits dont certains sont spécifiques aux « anchor » textes.

Les travaux basés sur une approche experte pour la traduction des EN sont rares. Nous pouvons citer comme exemple celui de (Gornostay & Skadiņa, 2009) qui permet la traduction des toponymes de l'anglais vers le letton. Ce travail se base sur l'utilisation d'un dictionnaire, d'une part, et sur l'utilisation des patrons syntaxiques, d'autre part. Ces patrons consistent à translittérer le toponyme ou le translittérer et le traduire ou le translittérer et lui ajouter une nomenclature.

En résumé, dans l'approche experte de la TA, les traductions sont construites en utilisant d'importants dictionnaires et des règles linguistiques sophistiquées. Les systèmes de TA basés sur cette approche fournissent un bon niveau de qualité de traduction dans des domaines non spécifiques. Leur adaptation à un domaine spécifique est possible mais elle représente un coût important en termes de temps et de moyens.

Quant à l'approche empirique, les traductions sont produites à partir d'événements rencontrés dans des corpus bilingues et à partir desquels le système va puiser ses connaissances. Pour fournir des traductions de bonne qualité, les systèmes de TA basés sur cette approche nécessitent une quantité importante de corpus bilingue. L'apprentissage des modèles dans cette approche est un processus rapide, automatique et peu coûteux.

Notons que d'autres systèmes profitent des avantages des deux approches en les combinant. Ces systèmes sont qualifiés d'hybride.

# Conclusion

Dans ce chapitre, nous avons commencé par discuter les définitions attribuées à la notion d'EN tout en montrant leurs limites. Cette discussion nous a permis de retenir une définition appropriée à notre travail. Ensuite, nous avons recensé les travaux effectués sur la catégorisation des EN qui est une partie intégrante du processus de reconnaissance

automatique. Ce recensement permet la proposition d'une hiérarchie de type relative à un domaine donné. Puis nous avons présenté les différentes approches de reconnaissance en précisant les avantages et les inconvénients de chacune d'elles. Les travaux présentés et basés sur ces approches peuvent nous aider à choisir les formalismes adéquats à notre tâche. Enfin, nous avons détaillé et comparé les différentes méthodes adoptées pour la traduction automatique. Les travaux de traduction présentés dans ce chapitre montrent que la traduction des EN utilisant une approche linguistique n'est pas bien développé et nécessite encore beaucoup d'efforts pour palier les lacunes existantes.

# Chapitre 2 : Typologie des EN arabes

La typologie des EN arabes proposée est basée sur l'étude effectuée sur les différents concepts et formes des EN identifiées dans le corpus d'étude que nous avons établi. Cette étude utilise la définition retenue caractérisant une EN valide et inspirée des définitions déjà présentées et discutées dans le chapitre précédent. L'étude effectuée sur les EN du domaine du sport, sur lequel va porter notre recherche, va nous permettre d'identifier les catégories nécessaires pour former une hiérarchie de type des EN et de recenser certains phénomènes linguistiques liés à leurs traitements. Notre choix du domaine du sport n'est pas arbitraire mais au contraire il est bien justifié parce qu'il constitue un domaine riche et qui peut recouvrir les différentes formes de dénomination les plus connues d'une EN.

Dans ce chapitre, nous commençons par présenter la définition que nous avons retenue pour la description d'une EN arabe valide. Ensuite, nous étudions les différents aspects et formes des EN arabes extraites du corpus d'étude appartenant au domaine du sport et en se basant sur la définition retenue. Après, nous présentons les catégories et les sous-catégories identifiées. Puis, nous donnons une hiérarchie de type des EN respectant la catégorisation proposée et le domaine choisi. Enfin, nous discutons les différents phénomènes linguistiques que nous pouvons rencontrer lors du traitement des EN arabes.

## 1. Définition retenue pour l'EN arabe

La définition que nous avons retenue est inspirée de plusieurs travaux (p.ex, (Friburger, 2002), (Tran, 2006), (Daille et al., 2000), (Sekine et al., 2002) et (Fourour & Morin, 2003)) et des encyclopédies libres (p.ex., l'Atalapédie et le Wikipédia). Cette définition consiste à considérer l'EN comme étant un nom propre dans son sens élargi (i.e., prénom, syntagme nominal), une expression numérique, une fonction, un événement ou un terme. En effet, il s'agit d'une catégorisation plus large et qui peut regrouper les différentes catégories de la langue arabe. De plus, toutes les définitions proposées pour les EN exigent le principe de référence et la majorité d'elles n'ont pas imposé l'existence d'autres critères. Dans ce cas, une EN n'est pas nécessairement unique et peut être un nom propre ou une expression de temps ou de quantité bien qu'il existe d'autres définitions qui exigent l'unicité. Donc, nous considérons que par exemple ملعب الطيب المهيري بصفاقس *mal`ab elTayib elmhyry biSafaaqus*

*(stade de taieb al mhiri à sfax)* est une EN mais ملعب الطيب المهيري *mal`ab elTayib elmhyry (stade de taieb al mhiri)* est aussi une EN. En effet, les deux entités réfèrent à un nom de lieu et exactement à un nom de stade.

# 2. Différentes formes de l'EN arabe extraites d'un corpus

Afin d'identifier les différentes formes de l'EN arabe, nous avons construit un corpus d'étude contenant des textes qui présentent des listes de noms officiels du domaine de sport disponibles sur Internet pour des pays arabes (p.ex., Tunisie, Algérie, Arabie Saoudite et Syrie) et des pays francophones (p.ex., France, Belgique et Canada) et du site www.kooora.com spécialisé dans le domaine du sport toutes les spécialités confondus. Ce site est intéressant notamment pour les noms des lieux car il couvre ceux des différentes spécialités sportives. Ainsi, le corpus d'étude contient 3000 textes et 1120000 mots qui couvrent 50000 noms de lieux sportifs. Le recensement des EN à partir de ce corpus est basé sur la définition d'une EN que nous avons retenue dans le paragraphe précédent. La liste des EN arabes recensées contient deux grandes classes : le nom propre et les entités numériques. Chaque classe est caractérisée par différentes formes et structures. Notons que les noms propres peuvent contenir des formes d'entités numériques. Dans ce qui suit, nous détaillerons ces deux classes.

## 2.1. Les noms propres

Rappelons que le nom propre est une sous-catégorie de nom, s'opposant au nom commun. Ainsi, un *nom propre* désigne toute substance distincte de l'espèce à laquelle elle appartient. Il ne possède en conséquence aucune définition spécifique, sinon référentielle, et n'a de signification qu'en contexte, subjective, ou par des éléments de sa composition.

Un nom propre appartient donc à un référent déterminé (une personne, un animal ou une chose), que ce référent, réel ou imaginaire, existe *naturellement* (un élément géographique par exemple) ou qu'il soit *artificiellement* créé par l'homme (une œuvre d'art, une œuvre littéraire, etc.).

Dans notre corpus, les noms propres peuvent eux-mêmes se subdiviser en trois sous formes : les noms de personnes, les noms de lieux et les noms d'organisations.

### 2.1.1. Les noms de personnes

Un nom de personne en arabe est composé au minimum par le prénom suivi par le nom de famille. A ces qualificatifs peuvent s'ajouter d'autres qualificatifs pour identifier d'une façon plus sûre une personne. C'est pourquoi le nom arabe peut être composé par différentes parties qui varient d'un pays à un autre et leur ordre n'est pas systématiquement observé. En effet, il n'existe pas de règles strictes stipulant la composition des noms de personnes en arabe.

Ainsi, un nom de personne peut contenir autre que le prénom et le nom de famille اللّقب *al-lakab*, le titre الصفة *elsifah* ex. اللاعب *ellaa`ib (le joueur)*, le surnom الكنية *alkunyah* ex. بيكاسو الكرة التونسية *bykaassuw alkura eltuwnisiya* pour désigner le joueur أسامة الدراجي *'ausaama eldarraajy Oussama al Daraji*, le patronyme (ou nom de filiation) النسب *elnasab ex.* محمد بن محمد نور بن آدم هوساوي *muHammad bin muHammad nur bin 'aadam huwsaawy (Mohamed fils de Mohamed Nour fils de Adam Houssaoui)*, le gentilé (ou nom d'origine) النسبة *elnisbah* ex. التونسي *eltuwnisy (le tunisien)*.

D'autres formes sont aussi utilisées comme le prénom d'un ancêtre, largement connu et utilisé comme une sorte de nom de famille, utilisant le nom آل *'aal* voulant dire *famille [de]*, par exemple آل سعود *'aal sa`uwd* comme dans خالد بن فيصل بن عبد العزيز آل سعود *khaalid bin faySal bin `abd al`aZyZ 'aal sa`uwd* (*Khaled fils de Fayçal fils de Abdelaziz* de la famille de Saoud), un moyen commode d'éviter d'avoir à décliner la filiation complète jusqu'à son arrière-arrière-arrière-arrière-arrière-arrière-grand-père Saoud, ce qui donnerait « Khaled fils de Fayçal fils de Abdelaziz fils de Abderrahman fils de Fayçal fils de Turki fils de Abdallah fils de Mohammed fils de Saoud »

Dans ce qui suit, nous détaillons chacun des qualificatifs cités auparavant.

#### Le surnom

Le surnom était d'un usage courant au temps de la Rome antique, il portait alors le nom de *cognomen*. Le surnom peut aussi s'ajouter au nom propre d'un individu et ainsi le distinguer, par rétronymie, de ceux qui s'appellent comme lui. Il peut avoir une valeur métaphorique, par exemple ابن بطوطة *'ibn baTTuwta* qui veut dire littéralement « l'enfant de canard » pour désigner quelqu'un de grand voyageur.

Le surnom peut être aussi honorifique ou de révérence et de vénération, par exemple أبو النور *'abuw elnuwr*, أبو الكرام *'abou alkiraam*.

Il existe une autre pratique dans le pays arabe de la région orientale qui consiste à nommer le père avec le prénom de son premier enfant en y ajoutant la particule أبو *'abuw* ou أم *'um* comme أبو عمار *Abou Ammar* ce qui veut dire littéralement « le père de Ammar » pour désigner le président Yasser Arafat.

Parfois le surnom est utilisé et connu plus que le prénom comme dans le cas de أبوبكر الصديق *'abuw bakr elSidyk* dont son véritable nom est عبد الله بن أبي قحافة التيمي *`abd allaah bin 'aby qaHaafah eltymy*.

De plus, le surnom peut s'inspirer des pratiques dialectales comme en Égypte où on donne le surnom de أبو علي *'abou `aly* à toute personne portant officiellement le prénom de حسن *Hasan*.

Le surnom peut être utilisé pour ridiculiser tel que أبوجهل *'abuw jahl*.

## Patronyme (Le nom de filiation)

Le mot « **patronyme** », d'origine grecque, signifie « nom du père ». Il peut désigner deux sortes de noms différents : le nom de famille, héréditaire, qui se transmet du géniteur ou du père à l'enfant, et qui reste en principe inchangé sur plusieurs générations ou peut également être, dans certaines cultures, le prénom du père d'une personne rappelé avec le prénom propre de cette personne (typiquement : X, **fils de Y** ou **ben Y**) ; il change donc à chaque génération. Dans notre contexte, nous suivons le deuxième cas comme par exemple عقبة بن نافع الفهري *`uqba bin naafa` alfihry*. Le mot بن *bin* (ou بنت *bint*) peut exister un certain nombre de fois pour bien qualifier la personne et préciser son identité tel que عقبة بن نافع بن عبد القيس الفهري *`uqba bin naafa` bin àbd alqays alfihry*.

Le prénom de la mère est rarement utilisé comme dans le cas du prophète عيسى `ysaa «Jésus» mentionné de nombreuses fois dans le Coran sous le nom de عيسى بن مريم *`ysaa bin maryam* «Jésus fils de Mariem».

Notons que le surnom peut être mentionné comme étant un nom de filiation comme par exemple عمر بن عبد الله بن أبي ربيعة المخزومي *ùmar bin `abd allaah bin 'aby raby`a almakhZuwmy*.

## Le prénom

Le prénom est la seule dénomination de l'identité intime de l'individu et il peut être soit simple (ex., يوسف yuwsuf) soit composé (ex., عبد الله *`abd allaah*). Le prénom peut porter le nom de dieu (ex., عبد القادر *`abd alqaadir*, عبد الرحمن *`abd elraHmaan*), ou un attribut du prophète (ex., أحمد *'aHmad*, المختار *almukhtaar*, الهاشمي *alhaashmy*) ou un nom du prophète (ex.,

إبراهيم *'ibraahym*, عيسى *`ysaa*, موسى *muwsaa*) ou d'occasion (ex., شعبان *sha`baan*, رمضان *ramadhaan*, العيد *al`yd*).

De plus, les arabes de confession chrétienne ont tendance à utiliser des prénoms s'inspirant de la Bible comme جون *juwn* (*Jean*) ou des prénoms d'origine européenne comme جورج *juwrj* (*George*).

La majorité des prénoms arabes sont des mots doués d'un sens par exemple نبيل *nabyl* qui veut dire «noble». Ce même mot peut être utilisé dans un texte comme un adjectif. D'autres prénoms peuvent être sous forme d'un verbe tel que يزيد *yazyd* qui veut dire *augmenter* conjugué à la troisième personne du singulier. Ces prénoms peuvent être une source de confusion dans le repérage des EN surtout avec l'absence des voyelles courtes. C'est pourquoi dans certains journaux le prénom d'une personne est écrit entre guillemets pour éviter cette ambigüité.

## Le gentilé

Le gentilé est le *nom d'habitants* par rapport à un lieu (pays, province, ville...). Il inclut entre autres : le lignage ou l'origine sociale de la personne par exemple التونسي *eltuwnisy* est celui originaire de la Tunisie et القرقني *alqarqny* qui est originaire de l'île de Kerkennah.

Le nom d'origine doit se terminer avec la lettre ي /Ya accompagnée de la voyelle courte /i/ vers la fin du mot, ce qui indique sémantiquement l'appartenance du prénom X à l'endroit Y. Il peut exprimer aussi l'affiliation à une religion et peut se retrouver sous forme d'adjectif exprimant le métier de la personne (ex., روحاني *ruwHaany*).

Notons qu'il existe des règles de formation du gentilé qui sont plus complexes et d'autres où le ي /Yaa est omise comme عطار *`aTTaar* pour désigner celui qui exerce le métier de vendeur ou مِعْطار *mi`Taar* pour désigner celui qui a du parfum.

## Le titre

Le titre peut être de noblesse comme الأمير *al'amyr le prince* ou honorifique, par exemple الشيخ *elshykh* ou même une référence à un métier comme اللاعب *ellaa`ib* (*le joueur)*.

Le **Tableau 3** récapitule les différentes possibilités pour la formation d'un nom d'une personne qui peuvent exister dans un texte.

| Titre | Surnom | Prénom | Nom de filiation | Nom de famille | Gentilé |
|---|---|---|---|---|---|
| اللاعب<br>*ellaa`ib* | بيكاسو الكرة التونسية<br>*bykassu alkura eltuwnisiyya* | أسامة<br>*'usaama* | *ابن المنجي*<br>*Ibn elmunjy* | الدراجي<br>*eldarraajy* | *التونسي*<br>*eltuwnisy* |
| اللاعب<br>*Al-laib* | بيكاسو الكرة التونسية<br>*Picasso alkura eltuwnisiyya* | أسامة<br>*'usaama* | | الدراجي<br>*eldarraajy* | *التونسي*<br>*eltuwnisy* |
| اللاعب<br>*Al-laib* | | أسامة<br>*'usaama* | | الدراجي<br>*eldarraajy* | *التونسي*<br>*eltuwnisy* |
| اللاعب<br>*Al-laib* | | أسامة<br>*'usaama* | | الدراجي<br>*eldarraajy* | |
| اللاعب<br>*Al-laib* | | | | الدراجي<br>*eldarraajy* | |
| | بيكاسو الكرة التونسية<br>*Picasso alkura eltuwnisiyya* | أسامة<br>*'usaama* | *ابن المنجي*<br>*Ibn elmunjy* | الدراجي<br>*eldarraajy* | *التونسي*<br>*eltuwnisy* |
| | بيكاسو الكرة التونسية<br>*Picasso alkura eltuwnisiyya* | أسامة<br>*'usaama* | *ابن المنجي*<br>*Ibn elmunjy* | الدراجي<br>*eldarraajy* | |
| | بيكاسو الكرة التونسية<br>*Picasso alkura eltuwnisiyya* | أسامة<br>*'usaama* | | الدراجي<br>*eldarraajy* | *التونسي*<br>*eltuwnisy* |
| | بيكاسو الكرة التونسية<br>*Picasso alkura eltuwnisiyya* | أسامة<br>*'usaama* | | الدراجي<br>*eldarraajy* | |
| | بيكاسو الكرة التونسية<br>*Picasso alkura eltuwnisiyya* | | | | |
| | | أسامة<br>*'usaama* | *ابن المنجي*<br>*Ibn elmunjy* | الدراجي<br>*eldarraajy* | *التونسي*<br>*eltuwnisy* |
| | | أسامة<br>*'usaama* | *ابن المنجي*<br>*Ibn elmunjy* | الدراجي<br>*eldarraajy* | |
| | | أسامة<br>*'usaama* | | الدراجي<br>*eldarraajy* | *التونسي*<br>*eltuwnisy* |
| | | أسامة<br>*'usaama* | | الدراجي<br>*eldarraajy* | |
| | | | | الدراجي<br>*eldarraajy* | |

**Tableau 3.** *Différentes variations du nom de personne أسامة الدراجي 'usaama eldarraajy*

Le **Tableau 3** montre les différentes variations de formation du nom de joueur أسامة الدراجي *'usaama eldarraajy*. Ces variations indiquent qu'il n'existe pas de règles strictes pour la formation d'un nom d'une personne. Ceci montre aussi qu'un même nom d'une personne peut exister sous différentes formes. La **Figure 6** illustre la répartition fréquentielle des formes d'un nom d'une personne.

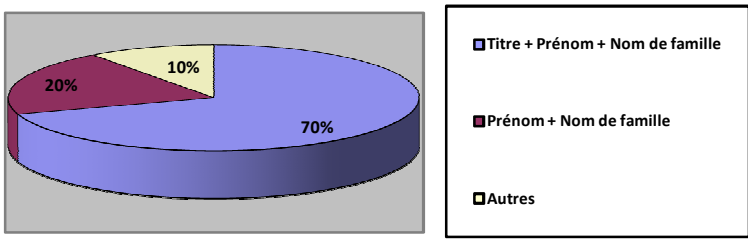


**Figure 6.** *Répartition des différentes formes d'un nom d'une personne*

Nous remarquons à travers cette répartition que la majorité des noms de personnes est formée par un titre suivi d'un prénom suivi d'un nom de famille (70%) ou d'un prénom suivi d'un nom de famille (20%). Les autres structures sont rarement utilisées dans notre corpus.

### 2.1.2. Les noms de lieux

Les noms de lieux en arabe, comme dans d'autres langues, désignent les villes, les pays, les villages, les montagnes et les fleuves. Dans notre corpus, cette catégorie inclut tout ce qui représente un lieu sportif tel que les noms de stade, de piscines, de salles de sport, de cité.

La liste des noms de lieux connus et existants dans le monde est relativement stable dans la mesure où les noms de lieux ne changent pas souvent. Toutefois, à l'instar des noms de personnes, certains noms de lieux sont ambigus. Par exemple, le mot *Tunisie* en arabe تونس *tuwnis* ou *Algérie* الجزائر *aljaZaa'ir* désigne le pays ou la capitale. De plus, le mot *Maroc* en arabe المغرب désigne soit le pays qui est le *Maroc* soit la grande région du Maghreb située en Afrique du Nord. Parfois et afin de lever l'ambiguïté, on appelle le *Maroc* المغرب الأقصى *almaghrab al'aqsaa* qui signifie le (Maroc lointain) et المغرب الكبير *almaghrab alkabyr* pour désigner la région entière du Maghreb.

Les exemples cités n'ont pas de répercussions majeures sur les performances d'un système de repérage des EN puisque l'ambiguïté reste tout de même entre deux catégories de noms de lieux. Par contre, la tâche se complique quand un nom de lieu dispose de plusieurs significations et qui n'ont aucun rapport les unes avec les autres. Par exemple, le mot *Yémen* en arabe اليمن *alyaman* peut être soit le pays, soit le nom commun signifiant la chance. En outre, la traduction du sens des noms des pays étrangers en équivalents arabes est une pratique courante. Cette pratique peut avoir un effet sur le repérage des EN. Par exemple, les Pays-Bas

الأراضي المنخفضة *al'araadhy almunkhafidha*, qui signifie «les basses terres», est utilisé en alternance avec le mot هولندا *houlanadaa*, qui reste le plus fréquemment employé, ou le Cap-Vert الرأس الأخضر *elra'as al'akhdhar* qui signifie littéralement « la tête verte».

Les noms de lieux spécifiques à notre domaine sont constitués généralement par d'autres types d'EN. Ils représentent des syntagmes nominaux formés par un mot introducteur (ou déclencheur) suivi par un autre type d'EN tel que les noms de personnes (p.ex., ملعب الطيب المهيري *mal`ab elTayib 'almhyry* (*stade de taeib el mhiri)*), les dates (p.ex., ملعب 14 جانفي *mal`ab 14 janfy (stade 14 janvier)*). Un nom de lieu sportif peut même contenir des EN de même catégorie par exemple ملعب رادس *mal`ab raadis (stade radès)*.

Par ailleurs, les noms de lieux sportifs étrangers se composent souvent d'un mot introducteur en arabe comme *stade* ملعب *mal`ab* et le reste des éléments de l'entité sont simplement sous forme d'une translittération simple du nom étranger ou même toute l'EN est translittérée comme dans Stade de la Libération (France) qui sera traduite en arabe en ستاد دو لا ليبرياسيون *staad duw lybiryaasyuwn* alors que la forme bien traduite on la trouve entre parenthèses si elle existe (ملعب التحرير *mal`ab eltaHryr (stade de la libération)*). Il arrive aussi qu'un nom de lieu sportif étranger se trouve en arabe bien traduit ; il suffit d'ajouter parfois le mot introducteur en arabe tel que Parc des Princes (France) qui existe dans un corpus arabe sous le nom ملعب حديقة الأمراء *mal`ab Hadyaqat al'umaraa*.

La **Figure 7** illustre la répartition des noms de lieu dans notre corpus.

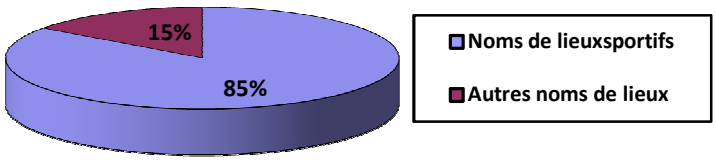


**Figure 7.** *Répartition des noms de lieux*

Comme indiqué dans la **Figure 7**, la majorité des noms de lieux dans notre corpus sont spécifiques au domaine du sport. En effet, les noms de lieux sportifs présentent 85% par rapport aux villes, aux pays, aux villages, aux montagnes et aux fleuves ou un nom quelconque de lieu. Cependant, il faut bien noter que les noms de lieux sportifs peuvent

contenir eux mêmes des noms de villes comme par exemple ملعب مدينة الباسل الرياضية بدرعا *mal`ab madynat albaasil elriyaaDiyya bidar`aa (stade de la cité sportive al bacel de Deraa)*.

### 2.1.3. Les noms d'organisations

Cette catégorie d'entités nommées inclut les noms des gouvernements, des clubs, des sélections et des fédérations. Les noms d'organisations sont assez nombreux et sont difficilement quantifiables puisque leur apparition et leur disparition dépendent de la situation dans le monde.

De plus, dans un texte, une même organisation peut changer entre l'usage d'une forme longue et d'une forme courte de son nom. Par exemple النادي الرياضي الصفاقسي *elnaady elriyaaDy elSafaaqusy* (*Club Sportif Sfaxien*) qui est une forme longue, peut exister dans un autre texte avec une forme plus réduite comme النادي الصفاقسي *elnaady elSafaaqusy* (*Club Sfaxien*).

Comme dans la plupart des langues, les noms d'organisations dans la langue arabe, peuvent être soit simples (un seul mot), soit complexes (deux mots ou plus). Pour les noms d'organisations sportives, ils sont souvent composés tels que الاتحاد التونسي لكرة القدم *al'ittiHaad eltuwnisy likurat alqadam (Fédération Tunisienne de Football)* connu sous le nom de FTF. Cependant, l'usage des acronymes en arabe est relativement faible si on le compare aux langues européennes comme le français et l'anglais. Ceci est dû au fait que la langue arabe n'a jamais adopté cette forme de réduction des noms propres. Les exemples trouvés dans notre corpus sont rares ; en fait, il s'agit plutôt d'acronymes étrangers. Dans ces exemples, on se contente souvent de transcrire phonétiquement l'acronyme comme dans le cas de FIFA qui s'écrit en arabe الفيفا *alfyfaa***.**

Par ailleurs, les noms d'organisations étrangères se composent souvent d'un mot introducteur en arabe comme *fédération* اتحاد *'ittiHaad*, *association* جمعية *jam`iya* et le reste des éléments de l'entité sont simplement sous forme d'une translittération simple du nom étranger comme dans Association CESAM qui sera traduite en arabe جمعية سيزام / فرنسا.

En observant la liste des noms d'organisations originaires des pays arabes, nous avons parfois constaté l'insertion de mots d'origine anglaise ou française dans la composition de l'entité. Par exemple شامبيون للأدوات والمعدات الرياضية *shaambyuwn lil'adawaat walmu`iddaat elriyaaDiyyah* (*champion pour les outils et les équipements sportifs)*, سيتي سبورت *syty sbuwrt (cité sport)* et سبورت للمعدات الرياضية *sbuwrt lilmu`iddaat elriyaDiyyah (sport pour les équipements sportifs)*. En outre, le nom d'une organisation peut être composé par d'autres EN

de même type ou non tels que برشلونة للأدوات الرياضية *barshaluwnah lil'adawaat elriyaDiyyah (Barcelone pour les outils sportifs)*.

Ainsi, nous pouvons déduire que les formes à reconnaître dans les noms d'organisations sont plus variées et plus complexes que celles des noms de personnes et de noms de lieux. En effet, les noms d'organisations en arabe sont souvent choisis d'une manière arbitraire et n'obéissent pas à des règles strictes ; n'importe quel mot peut théoriquement faire partie de cette catégorie. Ceci est dû surtout à l'irrégularité et à l'inconsistance dans l'usage alterné entre l'arabe écrit et les mots empruntés d'autres langues. Ce type de combinaison peut compliquer la tâche de repérage et de reconnaissance des EN.

La **Figure 8** illustre la répartition des noms d'organisations sportives dans notre corpus d'étude. Cette répartition est basée sur la structure de l'EN.

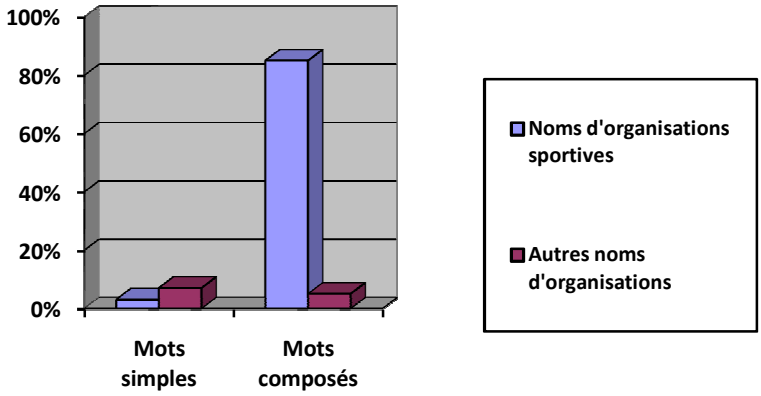


**Figure 8.** *Répartition des noms d'organisations sportives*

Comme indiqué dans la Figure 8, la majorité des organisations sportives sont composées (85%). En effet, les organisations sportives simples sont rares dans notre corpus tel que بروموسبور *bruwmuwsbuwr* (*promosport)*. On peut aussi remarquer l'existence d'autres EN de type organisation qui n'appartiennent pas au domaine choisi dont 5% sont simples et 7% sont composés.

### 2.1.4. Autres formes de noms propres

Dans notre corpus de sport, il existe d'autres formes d'EN qui sont fortement liées au domaine choisi qui est le sport. Il s'agit des EN de type *Nom de sport* qui est considéré comme un nom propre. Ce type inclut les noms de sports collectifs tels que كرة القدم *Kurat*

*alqadam (Football)* et les noms de sport individuels comme par exemple السباحة e*lsibaaHah (la natation)*.

### 2.1.5. Ambigüité des noms propres

La polysémie constitue l'un des problèmes majeurs dans un système de repérage des EN. En effet, les noms propres arabes sont la plupart du temps porteurs d'un sens particulier, il arrive souvent qu'ils présentent une ambiguïté que seul le contexte permet de résoudre. Par exemple, le nom de l'ex-président syrien حـافـظ الأســـد *Haafidh alasad* veut littéralement dire «le protecteur du lion». Ainsi même un locuteur natif méconnaissant cette personnalité politique peut tomber dans l'erreur et juger qu'il ne s'agit pas d'un nom propre, mais un syntagme nominal.

Afin de résoudre partiellement cette ambiguïté, il faut étudier le contexte immédiat (mot déclencheur) du nom propre et vérifier l'existence des marqueurs lexicaux qui peuvent précéder ou suivre ce dernier.

Le mot déclencheur peut se trouver à l'intérieur même du nom propre et peut prendre la forme d'un ou plusieurs mots ou d'une abréviation connue pour faire partie d'une EN comme par exemple dans l'EN ملعب الطيب المهيري *mal`ab elTayib 'almhyry* (*stade de Taib EL-Mhiri),* le mot déclencheur ملعب *mal`ab* (*stade)* appartient à l'EN à reconnaître et celui qui permet de déterminer sa catégorie.

## 2.2. Les entités numériques

Les entités numériques sont divisées en deux grandes formes d'ENA : les expressions de temps et les nombres.

### 2.2.1. Expression de temps

Les expressions de temps incluent les dates, la période et toute autre expression exprimant le temps. La plupart des entités temporelles dans la langue arabe sont identifiables grâce à une liste de marqueurs lexicaux comme par exemple *jour, mois, année*, etc. Concernant les dates, il existe une différence dans l'usage des calendriers qui varient d'un pays arabe à un autre tels que le calendrier grégorien (ex., *01 نوفمبر 2007*), le calendrier syriaque (ex., *02 تشرين الثاني 2007*) et le calendrier musulman (ex., *21) من شوال 1428 هـ*). Les tableaux 4, 5 et 6 donnent une idée sur les différents mois pour chaque type de calendrier déjà cité.

| N° | Nom du mois | Nombre de jours |
|---|---|---|
| 1 | muHarram محرم | 30 |
| 2 | Safar صفر | 29 |
| 3 | raby` al'awwal (Rabi I) ربيع الأول | 30 |
| 4 | Raby` elthaany (Rabi' II) ربيع الثاني | 29 |
| 5 | jumaadaa al'awwal l (Jumada I) جمادى الأول | 30 |
| 6 | jumaadaa elthaany (Jumada II) جمادى الثاني | 29 |
| 7 | rajab رجب | 30 |
| 8 | sha`baan شعبان | 29 |
| 9 | ramadhaan رمضان | 30 |
| 10 | shawwaal شوال | 29 |
| 11 | dhu alqi`dah ذو القعدة | 30 |
| 12 | Dhu alhijjah ذو الحجة | 29 ou 30 |

**Tableau 4.** *Mois du calendrier musulman*

| N° | Nom du mois | Nombre de jours |
|---|---|---|
| 1 | janvier جانفي | 31 |
| 2 | février فيفري | 28 ou 29 |
| 3 | mars مارس | 31 |
| 4 | avril أفريل | 30 |
| 5 | mai ماي | 31 |
| 6 | juin جوان | 30 |
| 7 | juillet جويلية | 31 |
| 8 | août أوت | 31 |
| 9 | septembre سبتمبر | 30 |
| 10 | octobre أكتوبر | 31 |
| 11 | novembre نوفمبر | 30 |
| 12 | décembre ديسمبر | 31 |

**Tableau 5.** *Mois du calendrier grégorien*

| N° | Nom du mois | Nombre de jours |
|---|---|---|
| 1 | كانون الثاني *kaanuwn elthaany* | 31 |
| 2 | شباط *shabbaat* | 28 ou 29 |
| 3 | آذار *'aadhaar* | 31 |
| 4 | نيسان *nysaan* | 30 |
| 5 | أيار *'aayaar* | 31 |

| | | |
|---|---|---|
| 6 | حزيران *Huzayraan* | 30 |
| 7 | تموز *tammuwZ* | 31 |
| 8 | آب *'aab* | 31 |
| 9 | أيلول *'ayluwl* | 30 |
| 10 | تشرين الأول *tashryn* al'awwal | 31 |
| 11 | تشرين الثاني *tashryn elthaany* | 30 |
| 12 | كانون الأول *kaanuwn al'awwal* | 31 |

**Tableau 6.** *Mois du calendrier syriaque*

Vu que notre corpus d'étude englobe des textes de différents pays arabes, nous avons identifié d'autres nominalisations pour les mois dépendant du pays. Par exemple, en Maroc les mois sont comme suit : يناير *yanaayir*, فبراير *fibraayir*, مارس *maaris*, أبريل *'abryl*, ماي *maay*, يونيو *yuwnyuw*, يوليوز *yuwlyuwZ*, غشت *ghisht*, شتنبر *shitanbar*, أكتوبر *'uhtuwbar*, نونبر *nuwnambar*, دجنبر *dijanbar*.

Les jours de la semaine dans la langue arabe sont habituellement précédés par le mot يوم qui signifie « jour ». De plus, l'article défini /Al/ doit être obligatoirement collé au nom du jour. Ces jours sont résumés dans le **Tableau 7**.

| **Jour** | **Arabe** | **Français** |
|---|---|---|
| يوم *yawm* (facultatif) | الأحد *al'a Had* | Dimanche |
| | الاثنين *al'itthnayn* | Lundi |
| | الثلاثاء *elthulaathaa* | Mardi |
| | الاربعاء *el'irbi`aa* | Mercredi |
| | الخميس *alkhamys* | Jeudi |
| | الجمعة *aljumu`ah* | Vendredi |
| | السبت *alsabt* | Samedi |

**Tableau 7.** *Jours de la semaine*

Dans notre corpus d'étude, nous avons identifié, avec un taux faible, l'existence des dates sous cette forme : Jour (facultatif) + 1-31(arabe ou indien) + mois grégorien (facultatif) + /ou mois solaire syriaque (facultatif) + année (arabe ou indien) + (facultatif) « correspondant » + 1-31 (indien) + mois hégirien + année hégirienne (indien) comme par exemple

الأحد 11 مارس/آذار2012 م الموافق 17 ربيع الثاني 1433 هـ

*al'aHad 11 maaris/'a adhaar 2012 m almuwafaq 17 Rabi` elthaany 1433 h*

Dimanche, 11 mars/Adhar 2012 miladi correspondant au 17 Rabi' al-thani 1433 hégirien

La plupart des dates identifiées, dans le corpus d'étude, étaient affichées en chiffres arabes et parfois en toutes lettres.

### 2.2.2. Nombre

Les nombres incluent principalement les systèmes de mesures (poids, distance, volume, vitesse), les pourcentages, ainsi que les devises.

Quand les unités de mesure sont employées dans un texte en arabe, l'usage des abréviations est systématique. Dans ce cas, on peut trouver des expressions comme : كغ 10 /*dix kilos* ط3 / *3 tonnes,* ص25/*25 cm*. Par exemple, la course de 100 mètres est transcrite en arabe par سباق مائة متر *sibaaq maat mitr.*

L'usage des pourcentages et des expressions monétaires se caractérise par l'emploi d'un signe particulier, le signe du pourcentage % dans le premier cas et les symboles monétaires pour les devises : $ pour le *dollar*, € pour l'*euro,* ¥ pour le *Yen*. Il arrive aussi que les symboles monétaires soient omis pour être remplacés par l'équivalent en arabe, qui est simplement une translittération du mot latin.

Les chiffres en arabe peuvent s'écrire aussi de différentes façons. Par exemple le chiffre « un » peut s'écrire 1 (chiffre arabe) ou I (chiffre romain) ou١ (chiffre indien).

Ainsi, l'usage dans la langue arabe des systèmes de mesures, des pourcentages ou des devises suit une convention et des règles d'écriture bien établies. Cette approche facilite la formulation des règles de repérage pour cette catégorie d'EN.

## 2.3. Autres formes d'EN arabes

Il existe d'autres catégories d'EN qui sont liés au domaine de sport. Ces catégories sont représentées sous forme de *Terme*, de *Fonction* et d'*Evénement*.

Le type *Terme* englobe tous les termes sportifs qui sont sous forme de règles, de technique comme par exemple ضربة مرمى *dharbat marmaae* (*touche)*, ضربة جزاء *dharbat jazaae (tir au but)*, de fautes telles que تسلل *tasallol* (*hors jeu)*, مخالفة *mukhaalafah* (*coup franc)* ou de punition comme ورقة حمراء *waraqah hamraae* (*carton rouge)*, ورقة صفراء *waraqa safraae (carton jaune)*.

Le type *Fonction* inclut les fonctions administratives telles que رئيس جامعة *ra'iyis jaam`i`a président de fédération* et رئيس جمعية *ra'iyis jam`iya* (*président d'association)* et les fonctions sportives telles que حكم *hakam (arbitre)* et مدرب *mudarrab (entraineur)*.

Le type Evénement représente les coupes continentales telles que كأس تونس لكرة القدم *k'as tuwnis likorat alqadam (Coupe Tunis de football)* et كأس الأمير فيصل بن فهد *kaas al'amyr faiySal bin fahd*

(*Coupe de prince Faiçal fils de Fahd)* et كأس ولي العهد *k'as waliy al`ahd (Coupe du Prince héritier)*, les coupes mondiales comme 2010 كأس العالم لكرة القدم *k'as al`alam likorat alqadam 2010* (*coupe du monde de football 2010)* ainsi que les jeux olympiques d'été par exemple الألعاب الأولمبية الصيفية 2008 *al'al`aab al'uwlambiyya elSayfiyya 2008* (*jeux olympiques d'été 2008)* ou d'hiver par exemple 2006 ألعاب أولمبية شتوية *al'al`baab al'uwlimbiyya elshitwiyya 2006* (*jeux olympiques d'hiver 2006)*.

# 3. Identification d'un modèle typologique des EN

Le modèle typologique des EN, que nous avons proposé, est dégagé à partir de l'étude des différentes formes des EN arabes que nous avons effectuée. La hiérarchie de ce modèle est inspirée des conférences MUC (Grishman, 1995). Celle-ci ne diffère pas beaucoup des autres hiérarchies dédiées pour les autres langues. En effet, les catégories composant la hiérarchie proposée sont communes pour les autres langues. Cependant, notre contribution se focalise essentiellement dans le raffinement effectué pour les différentes catégories à des niveaux différents. En effet, aux catégories des conférences MUC, nous avons ajouté les catégories suivantes : *Fonction*, *Terme* (ex., touche, corner, tir au but) et *Evénement* (ex., coupe du monde 2010). Outre la catégorie *Nom Propre* qui nous intéresse, nous avons dérivé celle de *Nom de lieu*. Cette dernière est raffinée par deux autres sous-catégories : *Emplacement* et *Ville*. La catégorie *Emplacement* est divisée en cinq sous-catégories : *Salle de Sport*, *Complexe*, *Cité Sportive*, *Stade* et *Piscine*.

La **Figure 9** illustre la hiérarchie proposée.

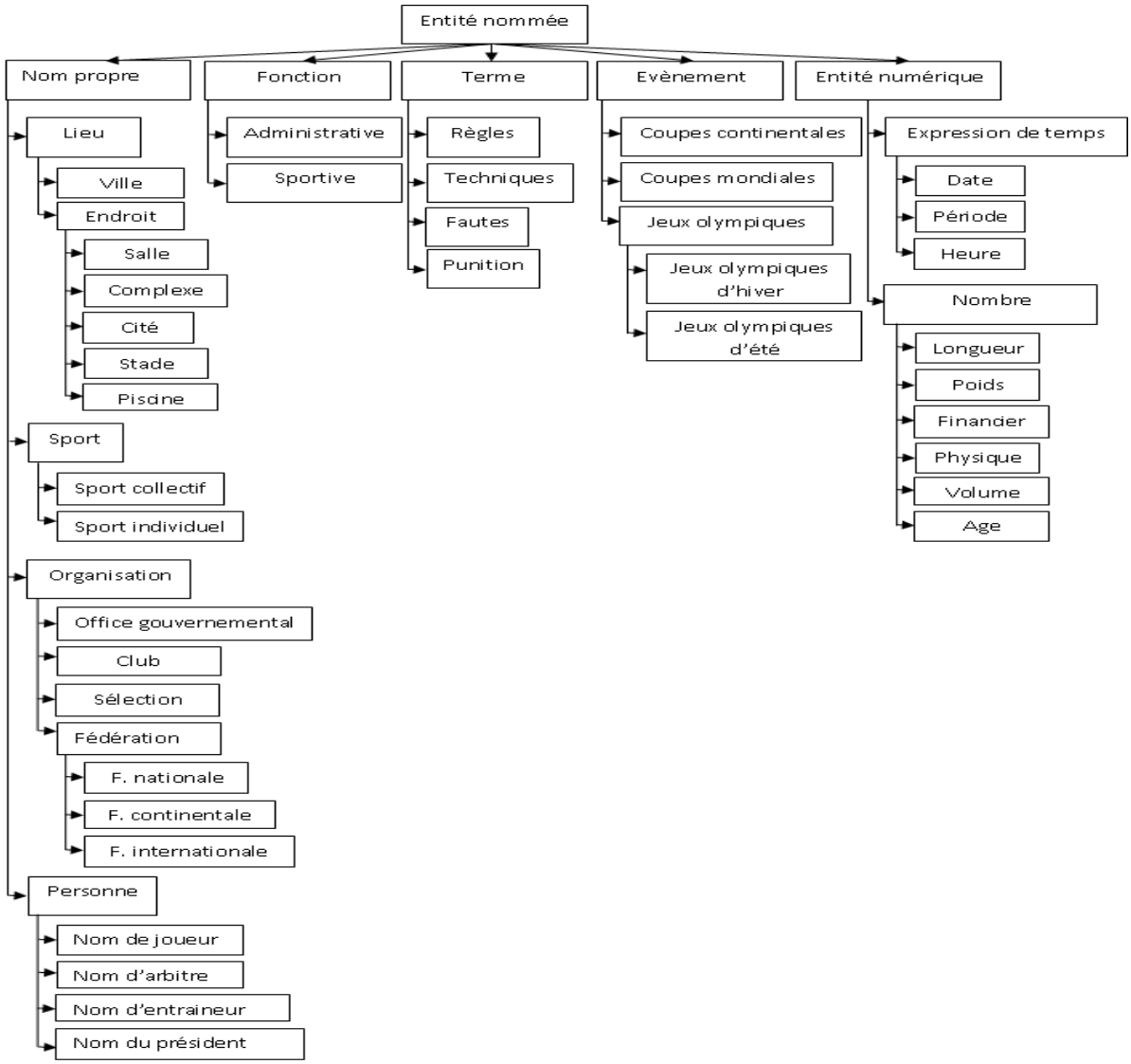


**Figure 9.** *Hiérarchie des EN*

La hiérarchie proposée nous permet le typage des principaux constituants d'une EN donnée à partir des catégories prédéfinies. En effet, une EN peut être composée par d'autres EN. Par exemple, L'EN ملعب الطيب المهيري *mal`ab elTayib almhyry (stade el Taeib el mhiri)* appartient à la catégorie *Stade* et elle contient aussi un nom composé الطيب المهيري *elTayib almhyry* de catégorie *Nom de Personne*. Notons que *Stade* et *Nom de Personne* sont deux sous-catégories de *Nom Propre*. Il est évident que le fait qu'une EN contient d'autres EN, peut engendrer des problèmes différents comme la polysémie et la métonymie (Poibeau, 2005). Cependant, cela prouve l'existence de la notion d'imbrication des EN.

# 4. Phénomènes linguistiques rencontrés

L'étude des différentes formes d'EN nous a montré l'existence de plusieurs phénomènes linguistiques qui peuvent causer des problèmes dans le processus de reconnaissance des EN arabes et même dans celui de traduction (Ben Hamadou et al., 2010). Parmi ces phénomènes, nous citons les suivants :

## 4.1. L'agglutination

La langue arabe est une langue fortement agglutinante du fait que les clitiques se collent aux substantifs, verbes, adjectifs auxquels ils se rapportent. De ce fait, nous trouvons des particules qui se collent aux radicaux en empêchant leurs détections, comme par exemple le mot «وبلعبه» / « et + par + jeu + sa » avec « و /et » indique une conjonction, «ب/par » présente une préposition, «لعب/ jeu» est un nom et «ه/ sa » désigne un pronom personnel. Ce qui rend son analyse automatique une tâche pénible à réaliser. D'ailleurs, contrairement aux langues romaines, son analyse ne nécessite pas seulement la vérification de l'appartenance de chaque mot du texte au dictionnaire et à la liste de formes fléchies et dérivées qui en découle mais aussi de donner tous les découpages potentiels en morphèmes.

Ces phénomènes d'agglutination augmentent, considérablement, le taux d'ambiguïté en introduisant d'autres supplémentaires au niveau de la segmentation des mots.

L'identification de ces particules nécessite un traitement spécifique surtout si le constituant est un nom composé (cas des toponymes) car la (les) particule(s) s'agglutine(nt) seulement à la première partie du nom composé. En effet, il est nécessaire de reconnaître la totalité du mot composé pour isoler la particule agglutinée (i.e., بـحمام الأنف (à Hammam lif)).

## 4.2. La détermination

Certains constituants du domaine du sport sont toujours déterminés (le cas des adjectifs). D'autres peuvent être déterminés ou non sans qu'il y ait des règles qui régissent ces différentes situations. Prenons le cas des toponymes : nom de ville qui suit directement la catégorie comme par exemple l'EN ملعب صفاقس (stade Sfax) où le toponyme est non déterminé et ملعب الرياض (stade du Ryadh) où le toponyme est déterminé.

Cette situation peut être aussi rencontrée pour les anthroponymes tels que ملعب شتيفي غراف (stade Steffy Graf) qui est indéterminé et ملعب الشاذلي زويتن (stade al Chadli Zouiten) qui est déterminé.

Le traitement de ce problème nécessite l'ajout dans le dictionnaire d'un trait supplémentaire sur la possibilité ou non d'une détermination.

## 4.3. Longueur des noms propres

Contrairement aux langues latines, les noms propres arabes ne commencent pas par des lettres majuscules (la majuscule n'existe pas dans la langue arabe). Les noms propres sont donc difficiles à identifier. Aussi, leur longueur n'est pas connue d'avance et peut dépendre des traditions de la région dans laquelle est née la personne. Ainsi, dans les noms de lieux sportifs, ils peuvent s'écrire sous forme d'un nom seul (ملعب الأسد, *stade Al-ASSAD* ) ou d'un nom et prénom (ملعب الطيب المهيري, *stade Ettaieb AL-MHIRI* ) ou d'un nom et prénom précédés d'un titre de noblesse (ستاد الملك عبد الله, *stade Roi Abdallah*) ou suivi de fils de (ولد) pour certaines régions (ستاد سحيم بن حمد, stade Sahim Bin Hamad). Par ailleurs, il n'est pas possible de mettre dans un dictionnaire tous les noms propres avec toutes leurs variantes d'écriture.

## 4.4. La syntaxe

Dans la langue arabe, la grammaire de construction des EN est riche et très variée. En effet, la longueur des EN (ou le nombre de constituants) ne peut pas être connue à l'avance ; elle est variable. Pour compléter le sens et le rendre non ambigu, on a tendance à ajouter un adjectif supplémentaire (i.e., municipal, olympique, national) tel que الملعب **الأولمبي** بالمنزه *Stade **olympique** d'Elmenzah* ou le nom de la ville ou le nom d'une ville suivi du nom du pays ملعب مدينة الباسل الرياضية **بدرعا** *Stade de la cité El-BACEL sportive à **Deraa*** etc. Ainsi, une EN peut contenir une partie essentielle et une autre facultative qui vient juste pour l'enrichir ou le rendre non ambiguë.

Notons aussi, qu'un même type de constituant peut se trouver à des positions différentes. Ce changement de position s'accompagne d'un changement de la structure de l'EN notamment au niveau des conjonctions et de la forme de détermination de certains constituants. C'est principalement le cas de l'adjectif qui ne suit pas toujours le nom auquel il se rapporte. Par exemple, dans l'EN ستاد عمان الدولي *Stade Amman international* l'adjectif *الدولي international* ne suit pas directement le nom avec lui il s'accorde ( *ستاد* stade) alors que dans l'EN الأستاد الوطني

في بانكوك *Stade national de Bangkok* l'adjectif *الوطني national* vient juste après le nom avec qui il se rapporte. On peut remarquer ici que contrairement au premier exemple le mot stade est déterminé et une préposition est ajoutée (في).

La position des toponymes est aussi variable telle que le cas **ستاد حلب الدولي** *stade Haleb international* où le toponyme حلب *Haleb* est au milieu de l'EN et le cas de l'EN **استاد الملك فهد الدولي بالرياض** *stade Roi Fahd international de Riadh* où le toponyme *الرياض Riadh* vient à la fin de l'EN.

A tous ces phénomènes, nous pouvons ajouter aussi les différentes forme d'écritures d'un même mot notamment les mots d'origine étrangère. Par exemple, le mot stade en arabe peut s'écrire ستاد *staade*, استاد *istaad*, إستاد *'istaad*.

De plus, nous remarquons l'existence des emprunts dans les prénoms. En effet, plusieurs noms de personnes trouvent leurs origines dans la langue perse (ex. Nermine نرمين = douce dans la langue perse). Cela est du vu l'importance de la distribution géographique des pays arabes et leurs liens avec les pays musulmans non arabophones.

En outre, nous constatons d'une part, les ambiguïtés des déclencheurs. Par exemple, le mot stade peut être un déclencheur pour des noms de stades comme « ملعب الطيب المهيري » (*stade de taeib el mhiri)* mais aussi pour certaines équipes comme «**الملعب التونسي**» (*Stade Tunisien)*. D'autre part, l'ambiguïté des noms de villes et de capitales. Par exemple, le mot «تونس, *tuwnis*» peut désigner la République Tunisienne ou Tunis, la capitale. C'est aussi le cas de «الجزائر, *aljaZaa'ir*» qui peut désigner l'Algérie ou l'Alger.

# Conclusion

Dans ce chapitre, nous avons présenté la définition de l'EN que nous avons retenue et qui nous semble la plus appropriée au contexte d'étude. Ensuite, nous avons effectué une étude sur les différentes formes de l'EN arabe. Cette étude nous a permis, en particulier, de dégager une hiérarchie des EN arabes dans le domaine du sport mais qui peut servir comme point de départ pour d'autres domaines, connaissant la richesse de ce domaine. Après, nous avons mis en avant les différents phénomènes linguistiques liés au repérage des EN et à leur traduction.

La hiérarchie proposée nous a permis le typage des principaux constituants d'une EN donnée à partir des catégories prédéfinies et de prévoir l'effet de la notion de l'imbrication des EN sur le processus de reconnaissance. Ainsi, une modélisation à travers un ensemble de traits peut

être une solution appropriée pour représenter explicitement cette notion. Cette modélisation fera l'objet du chapitre suivant.

# Chapitre 3 : Un modèle de représentation des EN arabes

Le modèle de représentation des EN, que nous proposons, est basé sur l'étude des différents constituants d'une EN. En effet, de cette étude nous avons constaté la complexité de la formation d'une EN. C'est pourquoi une représentation formelle des EN aide à clarifier les composantes et à bien représenter la notion d'imbrication des EN. De plus, elle permet de rendre plus fiable la constitution de ressources linguistiques. Une telle modélisation peut représenter tous les constituants d'une EN arabe dans une forme standard et peut limiter l'impact des spécificités linguistiques. Cependant, l'élaboration d'une représentation formelle et générique d'une EN n'est pas une tâche facile et cela pour plusieurs raisons : d'une part, cette représentation doit prendre en compte la notion de récursivité et de longueur variable des EN. D'autre part, la représentation à proposer doit contenir aussi un nombre suffisant de traits capables de représenter n'importe quelle EN indépendamment du domaine et de la catégorie grammaticale des constituants. Autrement dit, les mêmes traits doivent satisfaire tous les types d'EN.

Dans ce chapitre, nous commençons par présenter le modèle de représentation formelle des EN. Cette présentation décrit les sources d'inspiration de ce modèle, sa structure, ses différents traits et principes. Ensuite, nous illustrons chaque principe de bonne formation d'une EN à travers des exemples. Enfin, nous étudions les différentes caractéristiques de ce modèle en montrant son indépendance vis-à-vis de la catégorie, du domaine et de la langue.

## 1. Aperçu sur la structure attribut-valeur

La structure attribut-valeur (SAV) est l'ensemble de paires trait-valeur qui représente un objet linguistique. Les SAV ont été utilisées principalement en syntaxe, où un objet est considéré comme étant une catégorie grammaticale : un nom, un SN, une phrase, etc. Mais elles sont utilisées aussi en morphologie et en phonologie et à priori on peut les utiliser pour représenter n'importe quel objet linguistique (y compris les règles, comme le fait HPSG). La **Figure 10** donne un exemple d'une SAV.

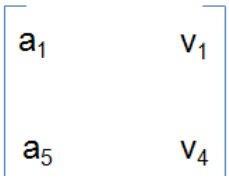


**Figure 10.** *Exemple d'une SAV avec des valeurs élémentaires*

Les valeurs des traits de la **Figure 10** sont simples. Mais, la valeur d'un trait dans une SAV peut être à son tour une SAV. Par exemple, dans la SAV de la **Figure 11** la valeur du trait $a_2$ est la SAV [$a_3$ $v_2$, $a_4$ $v_3$]. Pour cette raison, ces traits sont dits "à valeur catégorielle" et ils introduisent l'emboitement dans les SAV.

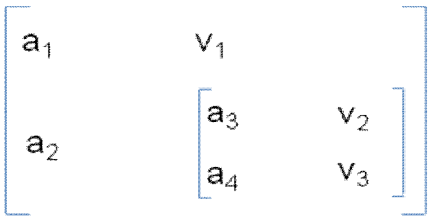


**Figure 11.** *Exemple d'une SAV avec une valeur complexe*

Dans le formalisme de SAV, plusieurs opérations peuvent être introduites telles que la négation et l'unification. L'opération de la négation est parfois utile pour décrire de façon succincte une longue disjonction de conditions. Concernant l'opération de l'unification, elle consiste à comparer et à combiner des SAV. Unifier deux catégories consiste à créer une troisième catégorie qui contient tous les traits de chacune des deux premières; si c'est impossible (parce que des traits sont incompatibles) alors on dit que l'unification échoue. Par exemple l'unification de la SAV de la **Figure 10** et celle de la **Figure 11** donne comme résultat la SAV représentée en **Figure 12**.

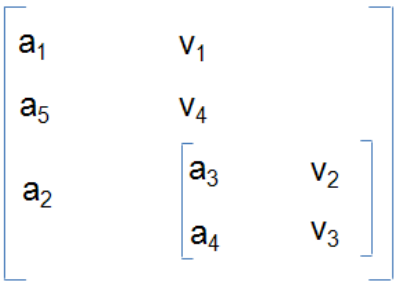


**Figure 12.** *Unification de deux SAV*

L'opération d'unification possède plusieurs propriétés intéressantes du point de vue d'un traitement automatique. C'est une opération commutative et associative. Dans le même contexte, on peut définir une relation entre les SAV ce qu'on appelle la subsomption. En effet,

une SAV A subsume une SAV B si B contient au moins autant d'informations que A (et peut en contenir plus). Par exemple si nous prenons la liste des SAV représentées dans la **Figure 13**, nous pouvons dire que A subsume B et C et que B subsume C alors que D ne subsume aucune autre SAV (et n'est subsumée par aucune).

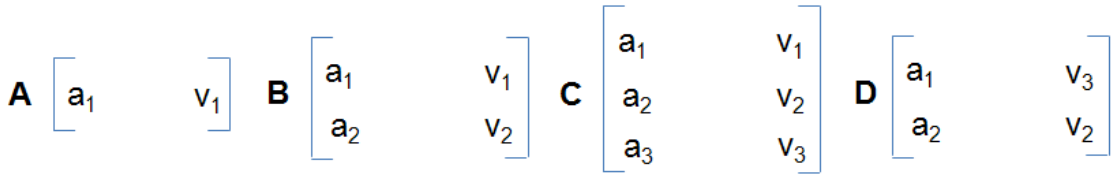


**Figure 13.** *Liste des SAV*

La subsomption définit un ordre (<) entre les SAV, dans lequel A<B si A subsume B. Dans la **Figure 13**, A<B<C mais D n'est comparable à aucune SAV.

L'emboitement dans la structure attribut-valeur peut être exploité dans la représentation des EN. En effet, une EN peut contenir d'autres EN. De plus, nous pouvons appliquer des opérations sur les EN telles que l'unification.

Parmi les formalismes qui reposent sur le modèle formel de la logique attribut-valeur, nous citons le formalisme des grammaires syntagmatiques guidées par les têtes HPSG (Head-driven Phrase Structure Grammar, HPSG) (Pollard & Sag, 1987) (Pollard & Sag, 1994), GPSG (Generalised Phrase Structure Grammar) (Gerald et al., 1985), CG (Categorial Grammar) (Morril, 1995) et LFG (Lexical Functional Grammar) (Bresnan, 2001).

# 2. Aperçu sur la notion Tête/Expansion

La notion de Tête/Expansion a été introduite en 1994 dans le système d'extraction terminologique LEXTER (Logiciel d'EXtraction TERminologique). Elle consiste à décomposer un syntagme en deux constituants : Tête (notée T) et Expansion (notée E). Cette décomposition est récursive dans le cas où un constituant est complexe. La **Figure 14** est un exemple de la procédure de décomposition du syntagme nominal "traduction automatique des langues naturelles".

Traduction automatique des langues naturelles
T : traduction automatique
    T : traduction
    E : automatique
E : langues naturelles
    T : langues

| E : naturelles |
|---|

**Figure 14.** *Exemple de décomposition*

Dans la **Figure 14**, on distingue deux syntagmes : "traduction automatique" et "langues naturelles". Le principe d'extraction de syntagmes consiste à utiliser une procédure de construction «ascendante». Dans cette procédure, on commence par les mots les plus «bas» dans la structure de dépendance. Ensuite, on construit le syntagme associé à un noyau dès que les (éventuels) syntagmes associés à ses compléments ont été construits. Enfin, chaque syntagme construit est composé :

- d'une **Tête**, correspondant au mot noyau
- d'une ou plusieurs **Expansions**, correspondant aux mots compléments, qui sont elles-mêmes des mots ou des syntagmes

L'objectif de ce traitement est la construction d'un réseau terminologique. Dans ce réseau, chaque syntagme est relié à sa Tête et à ses Expansions et à tous les termes dont il soit Tête, soit Expansion. On peut avoir aussi des syntagmes qui partagent la même Tête ou la même Expansion. La **Figure 15** illustre l'environnement terminologique du terme «traduction automatique».

| traduction automatique |
|---|
| **Tête de :** |
| traduction automatique des EN |
| traduction automatique des pages web |
| traduction automatique de site |
| **Expansion de :** |
| processus de traduction automatique |
| approches de traduction automatique |

**Figure 15.** *Environnement terminologique d'un terme*

Dans le même contexte, une EN peut être considérée comme un syntagme nominal qui peut avoir une Tête et une ou plusieurs Expansions. Par exemple, dans l'EN «Université de Sfax», la Tête est "Université" et l'Expansion est "Sfax". De même, on peut construire un réseau terminologique à cette EN. En effet, le mot «Université» peut être une Tête de «Sfax, Tunis, Franche-Comté, etc.» et une Expansion de «président, conseil, etc.».

# 3. Modèle de représentation formelle des EN arabes

Le modèle que nous proposons est utilisé pour formaliser et identifier les EN arabes. Ce modèle est inspiré des deux formalismes cités précédemment. En effet, sa structure est inspirée de celle « attribut-valeur ». La différence concerne essentiellement les attributs qui ne sont pas des catégories grammaticales. Ces attributs sont inspirés du concept «Head and Expansion» introduit par (Bourigault, 2002). A ces attributs, nous ajoutons d'autres qui permettent de bien caractériser une EN tels que son type. Les caractéristiques essentielles du modèle proposé sont :

- l'élément de la structure peut être atomique ou complexe,
- la structure interne d'un élément est définie par ses attributs et ses valeurs.

Dans ce qui suit, nous allons décrire la structure et les traits du modèle proposé.

## 3.1. Structure et traits proposés

Une EN dans notre modèle est définie d'une manière récursive. En effet, chaque EN possède un type «Type_EN» et est composée de deux parties : une essentielle et une autre extensionnelle. La partie essentielle est aussi une EN, elle admet donc à son tour une partie essentielle et une autre extensionnelle. Le type d'une EN est indiqué dans la plupart de temps par un déclencheur qui peut être un mot simple ou composé. Dans le modèle de représentation proposé, la partie essentielle est décrite par le trait «Tête_EN» et le mot déclencheur par le trait «Mot_declencheur». La partie extensionnelle représente la partie facultative qui compose l'EN. Elle n'admet pas généralement de type. Elle est précédée par un terme lexical «Elément_EN» (préposition, caractère spécial, etc.). C'est pourquoi, elle ne peut pas être considérée comme étant une EN mais elle peut contenir d'autres EN. Son existence ou non n'influence pas sur le sens de l'EN. Cette partie est décrite par le trait «Fin_EN».

Le modèle proposé respecte une grammaire bien déterminée. Cette grammaire est définie comme suit :

N = {EN, DECLENCHEUR, TETE, FIN, TYPE}, alphabet non terminal

Σ = {catégorie, entité, déclencheur, في, ب, -, [,], (,)}, alphabet terminal

Où :

- *catégorie* est l'ensemble des catégories identifiées dans la hiérarchie de type détaillée dans le chapitre 2
- *déclencheur* est l'ensemble des déclencheurs d'un domaine.
- *entité* est l'ensemble de toutes les EN d'un domaine.

**Règles de production**

L'ensemble de règles de production est formé par :

EN ⟶ DECLENCHEUR + TETE + FIN + TYPE

FIN ⟶ELEMENT + TETE + FIN

FIN ⟶ ε | adjectif

ELEMENT ⟶ PARTICULE | CARACTERE

PARTICULE ⟶ في | ب | ل

CARACTERE ⟶ - | : | [ | ] | ( | )

TETE ⟶ DECLENCHEUR +TETE + FIN + TYPE

TETE ⟶ entité | DECLENCHEUR

DECLENCHEUR ⟶ déclencheur | ε

TYPE ⟶ catégorie | ε

Le squelette du modèle proposé est alors présenté comme suit :

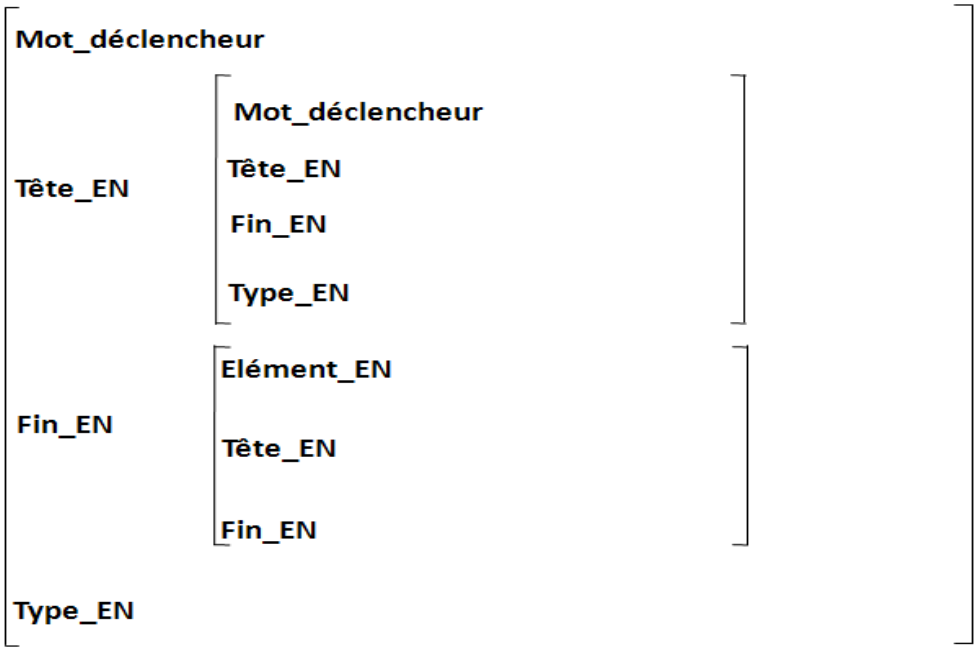


**Figure 16.** *Squelette d'une EN*

Comme indiquée dans la **Figure 16**, la valeur du trait «Tête_EN» peut être atomique ou structurée. Si elle est structurée, elle est composée alors par les traits : «Mot_déclencheur», «Tête_EN», «Fin_EN» et «Type_EN». La valeur du « Mot_déclencheur » peut être simple ou composée. En effet, un mot déclencheur peut être formé par un ou plusieurs mots. Sa valeur peut être aussi vide. La valeur de «Fin_EN» peut être atomique ou structurée. Si elle est structurée, elle sera composée par les traits : «Elément_EN», «Tête_EN» et «Fin_EN». Elle peut aussi être vide. La valeur du trait «Type_EN» est dans la plupart des cas simple ou composée. Elle est vide si la valeur du trait «Tête_EN» ne peut pas être considérée comme une EN. Cette valeur est égale à l'une des catégories définies dans la hiérarchie de type des EN. La valeur du trait «Elément_EN» est toujours simple.

Notons que les structures de traits de notre modèle de représentation ont pour principal objectif de décrire chaque EN composant l'EN principale et non pas de décrire des phénomènes linguistiques comme il existe dans la littérature. C'est le cas, par exemple, des traits utilisés par (Poibeau, 2005) qui décrivent le contexte dans lequel se trouve une EN, pour mettre en valeur les différents sens qui peuvent être affectés à une même EN.

Le modèle de représentation est doté de principes de bonne formation des représentations des EN. Ainsi, nous pouvons distinguer entre les structures bien formées et celles mal formées.

## 3.2. Principes de bonne formation

Dans le modèle de représentation proposé, trois principes doivent être satisfaits et sont utiles dans la phase de reconnaissance. Ces principes sont utilisés pour indiquer si une EN est bien formée ou non.

**Principe de saturation.** Une structure est dite *saturée* si elle représente une EN valide. Autrement dit, si le(s) trait(s) «Tête_EN» qui la compose(nt) sont tous non vides.

La **Figure 17** décrit un exemple d'une représentation formelle qui satisfait le principe de saturation.


**Figure 17.** *Structure d'une EN saturée*

Dans la **Figure 17**, la valeur du trait «Tête_EN» est élémentaire ; elle est non vide. De plus, le trait Type_EN est différent du vide. Donc, le mot الرياض *Riadh* est considéré comme étant une EN dont le type est *Ville*.

La **Figure 18** décrit un autre exemple avec plus qu'une entête et qui satisfait le principe de saturation d'une EN.

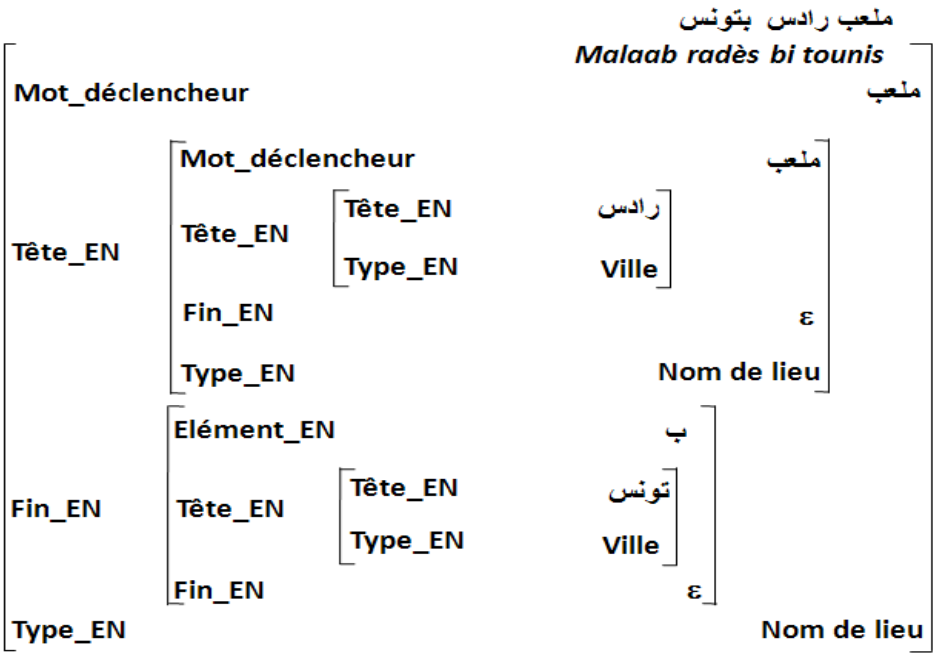


**Figure 18.** *Structure d'une EN saturée avec plus qu'une Tête_EN*

Dans la **Figure 18**, toutes les valeurs du trait «Tête_EN» sont non vides : ou bien structurées ou bien élémentaires. Donc, le syntagme ملعب رادس بتونس *mal`ab raadis bituwnis (stade radès en tunisie)* est considéré comme étant une EN dont le type est *nom de lieu*. Aussi, toutes les valeurs du trait Type_EN sont non vides. Donc, chaque valeur du trait Tête_EN est considérée comme une EN. Dans cet exemple, les EN sont au nombre de quatre (nombre de traits Type_EN dont la valeur est non vide) : ملعب رادس بتونس *mal`ab raadis bituwnis (stade radès en tunisie)*, ملعب رادس *mal`ab raadis (stade radès)*, رادس *raadis (radès)* et تونس *tuwnis (tunisie)*.

La **Figure 19** décrit un troisième exemple d'une EN avec plus qu'une entête et qui satisfait le principe de saturation d'une EN et qui admet une valeur du trait « Tête_EN » égale à la valeur du trait « Mot_déclencheur ».

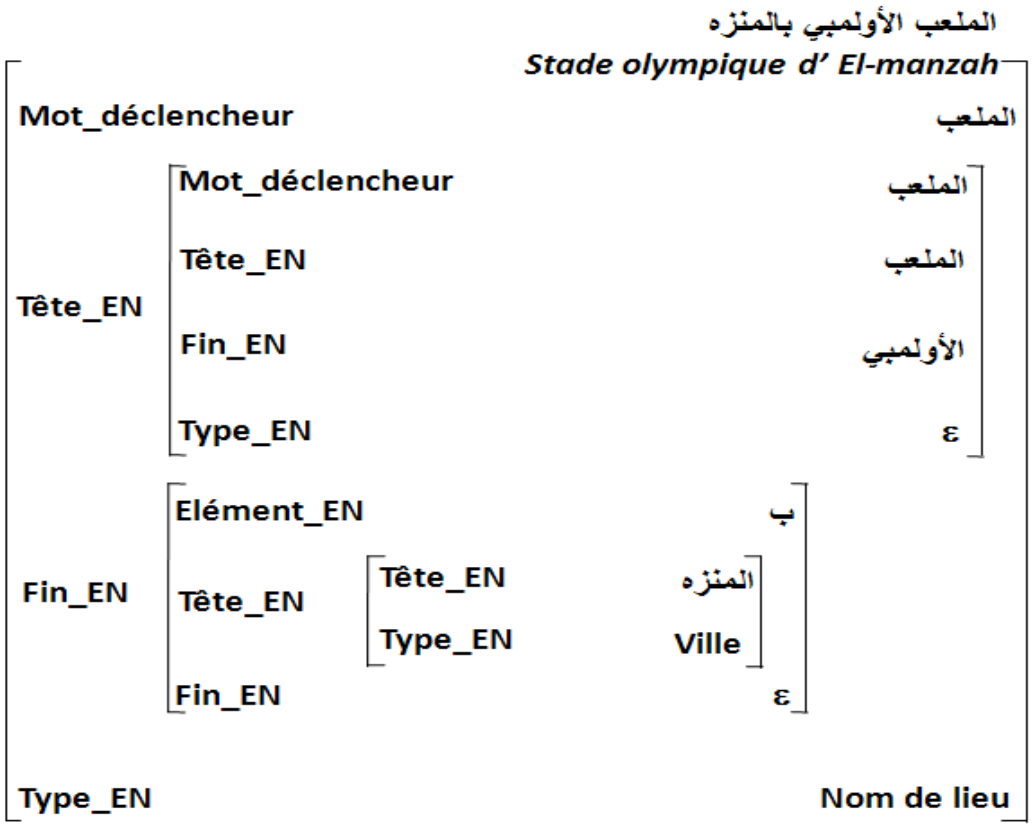


**Figure 19.** *Structure de l'EN « الملعب الأولمبي بالمنزه al-malaab al-oulimpii bil manzeh »*

Dans la **Figure 19**, toutes les valeurs du trait «Tête_EN» sont non vides : ou bien structurées ou bien élémentaires. Donc, le syntagme *الملعب الأولمبي بالمنزه almal`ab al'uwlamby bilmanzah stade olympique d'El-manazah* est considéré comme étant une EN dont le type est *nom de lieu*. Dans cet exemple, l'EN principale contient une autre EN : c'est celle dont la valeur du trait Type_EN est non vide : *المنزه almanazah*. L'entité *الملعب الأولمبي almal`ab al'uwlimby stade olympique* n'est pas considéré comme une EN car la valeur du trait Type_EN est vide.

Il est à noter que dans la structure d'une EN saturée, le nombre de trait Type_EN, dont la valeur est non vide, représente le nombre d'imbrication dans l'EN principale.

**Principe de non saturation.** Une structure est dite *non saturée* si elle peut être complétée pour qu'elle soit une EN. Autrement dit, lorsqu'elle est formée uniquement du trait «Fin_EN» ou si la valeur de son trait «Tête_EN» est vide.

Par exemple, dans le groupement de mots بتونس *bituwnis,* la valeur du trait «Tête_EN» est vide car ce groupement n'admet pas de type (prédéfinie dans la hiérarchie de type). Par conséquent, ce groupement n'est pas considéré comme étant une EN. Il ne satisfait pas le principe de saturation malgré qu'il contienne le mot تونس *tuwnis* qui est une EN. La **Figure 20** donne un exemple d'un modèle incomplet qui ne satisfait pas le principe de saturation.

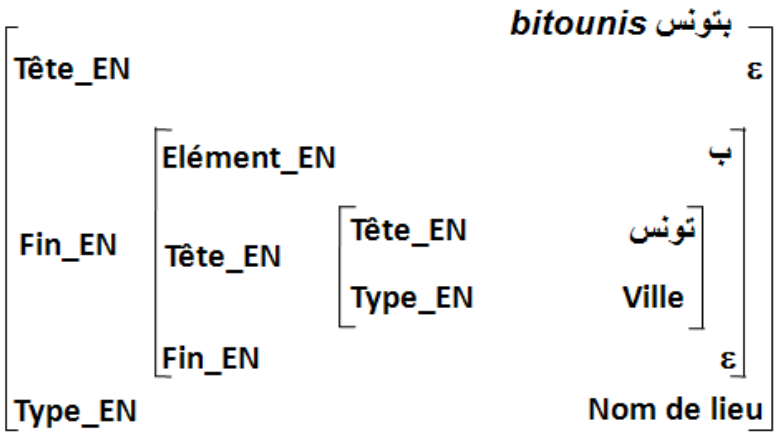


**Figure 20.** *Structure d'un modèle non complet*

L'exemple de la **Figure 20** peut être complété par d'autres mots pour que la structure de son modèle de représentation soit saturée et nous pouvons parler dans ce cas d'une EN. Par exemple, il est possible d'ajouter avant le mot بتونس *bituwnis* le groupement ملعب رادس *mal`ab raadis*. Ainsi, nous obtenons la même EN de la **Figure 18** qui satisfait le principe de saturation.

**Principe de propagation.** Le principe de *propagation* consiste en l'héritage de la même valeur du trait «Mot_déclencheur» dans des structures imbriquées successives tant que la valeur du trait «Type_EN» est la même. Ce principe est vérifié dans les exemples 18 et 19.

Les trois principes, ci-dessus, mentionnés nous permettent de se prononcer sur la bonne formation d'une EN. Cependant, ces trois principes ne sont utiles que lorsque le modèle est complété par des instances respectant la grammaire déjà définie.

## 3.3. Unification de deux représentations formelles

L'opération d'unification, que nous avons décrite au premier paragraphe, peut être appliquée à notre modèle. L'objectif est d'enrichir les structures conçues et de détecter l'incompatibilité et la propagation de traits.

Par exemple, la représentation formelle de l'EN ملعب رادس بتونس *mal`ab raadis bituwnis stade radès de tunis* donnée en **Figure 18** peut être unifiée avec la représentation formelle de l'EN décrite dans la **Figure 21**.

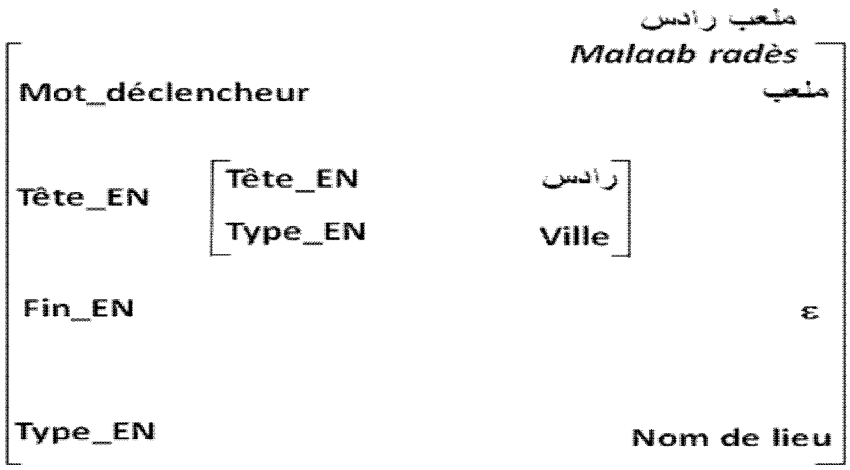


**Figure 21.** *Structure de l'EN malaab raadis bi tounis*

Le résultat de l'unification donne la même structure donnée par la **Figure 18**. Cela indique que la structure de la **Figure 21** subsume la structure de la **Figure 18**. En effet, la structure de l'EN *ملعب رادس بتونس mal`ab raadis bituwnis (stade radès en tunisie)* contient plus d'informations que la structure de l'EN *ملعب رادس mal`ab raadis (stade radès)*. Ainsi, nous pouvons dire que l'unification de deux représentations formelles nous permet d'avoir plus de détails sur l'EN et d'aller vers l'unicité de l'EN. D'après l'exemple que nous avons pris, nous pouvons déduire que *ملعب رادس stade radès* se situe à tunis.

## 3.4. Exemple illustratif

Dans cette section, nous donnons un exemple qui explique comment construire la représentation d'une EN. Pour cela, nous proposons un exemple d'EN qui a une structure relativement riche. Soit : *ملعب الملك عبد العزيزالدولي بالرياض mal`ab almalik `abd alaZyZ elduwaly bilriyaaD stade international roi Fahd à Riadh*. Cette EN est décrite par les règles suivantes :

EN → Mot déclencheur + Fonction + Prénom + Adjectif + Toponyme

Mot déclencheur → ملعب *mal`ab*

Fonction → الملك *almalik*

Prénom → عبد العزيز *`abd alaZyZ*

Adjectif →الدولي *elduwaly*

Toponyme →الرياض *elriyaD*

Dans l'EN ملعب الملك عبد العزيز الدولي بالرياض *mal`ab almalik `abd al`aZyZ elduwaly bilriyaaD*, le mot بالرياض *bilriyaaD* a pour fonction complément de lieu. Il vient juste pour enrichir l'EN. Par conséquent, son élimination n'affecte pas le sens de cette EN. C'est pourquoi, il va être

décrit par le trait «Fin_EN» comme suit : «Elément_EN» égale à ب *bi* et «Tête_EN» égale à الرياض *elriyaaD*. Notons que le mot الرياض *elriyaaD*, sans la particule ب *bi*, a son propre type. Pour cette raison, il sera une simple valeur du trait «Tête_EN». Aussi, le reste de l'EN ملعب الملك عبد العزيز الدولي *mal`ab almalik `abd al`aZyZ elduwaly* a à son tour son propre type. Il sera alors décrit par le trait «Tête_EN» qui va être structuré car cette EN contient des mots complémentaires. Dans cette EN, le mot complémentaire est l'adjectif الدولي *elduwaly* qui a le même rôle que le complément de lieu بالرياض *bilriyaaD*. Ainsi, il sera alors décrit par le trait «Fin_EN». De plus, le mot ملعب *mal`ab* joue le rôle d'un mot déclencheur. Ce qui reste est l'EN الملك عبد العزيز *almalik `abd al`aZyZ* dont le type est *Nom de personne*. Dans cette EN, le mot الملك *almalik* joue aussi le rôle d'un mot déclencheur. Cependant, l'EN عبد العزيز *`abd al`aZyZ* a pour type *Prénom*. Elle sera représentée alors par le trait «Tête_EN» comme valeur simple car عبد العزيز *Abd el Aziz* ne contient aucun mot complémentaire.

Ainsi, l'EN mentionnée est représentée dans notre modèle comme suit :

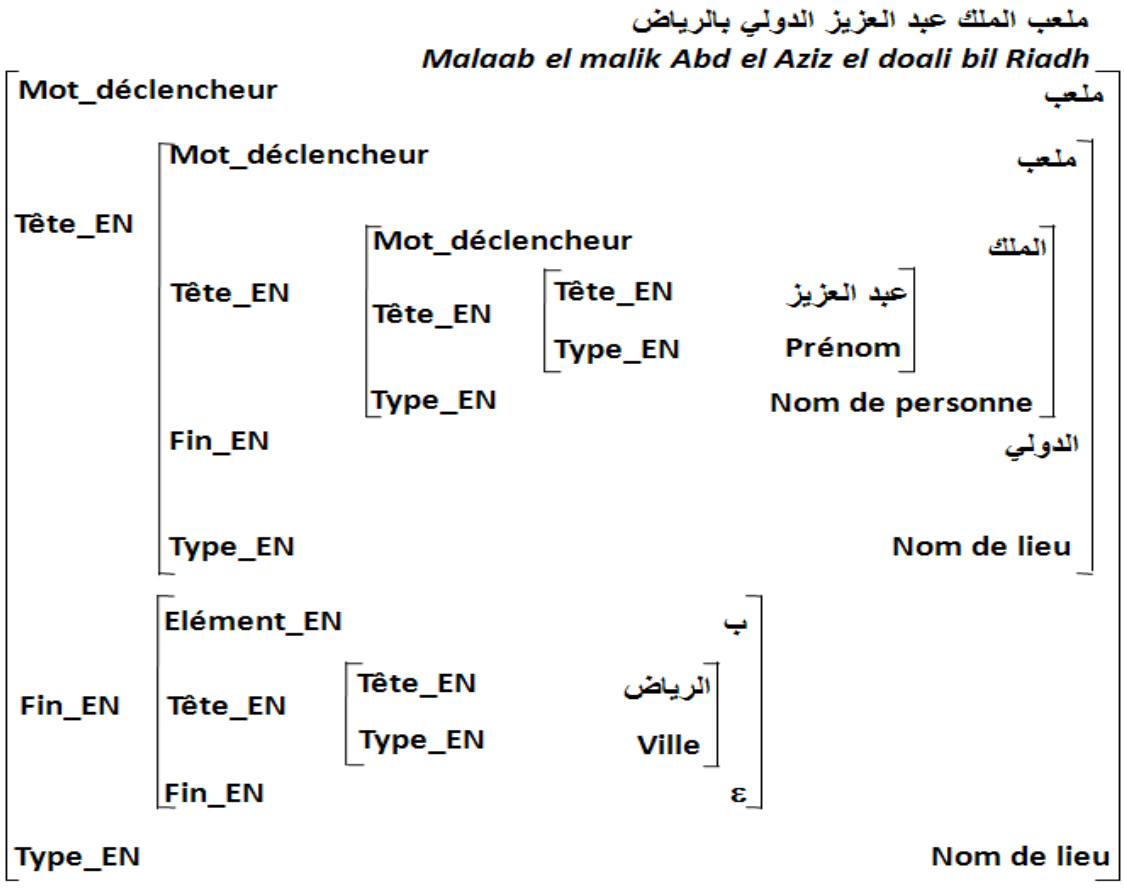


**Figure 22.** *Représentation formelle de l'entité «ملعب الملك عبد العزيز الدولي بالرياض»*

Dans la **Figure 22**, le principe de saturation est satisfait. En effet, toutes les valeurs du trait «Tête_EN» sont non vide. De plus, la valeur de chaque trait respecte le vocabulaire du domaine choisi.

## 3.5. Indépendance du modèle vis-à-vis du domaine

La représentation proposée est applicable indépendamment du domaine. En effet, ayant effectué une étude sur les noms de lieux, nous avons remarqué que toutes les EN de cette catégorie ont la même structure quelque soit le domaine. La **Figure 23** est un exemple d'une EN appartenant au domaine médical.

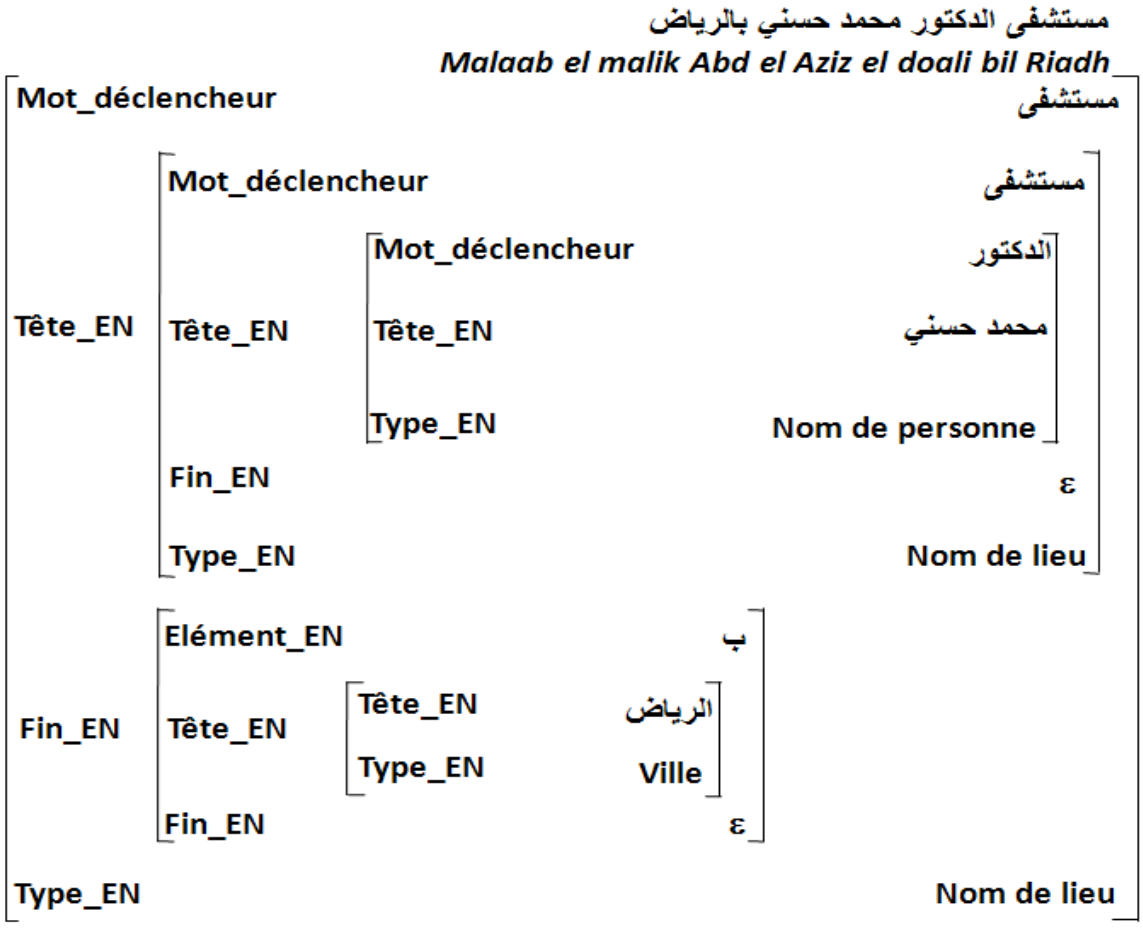


**Figure 23.** *La représentation formelle d'une EN appartenant au domaine médical*

Dans l'exemple de la **Figure 23**, l'EN *Hôpital du docteur Mohamed Hosni à Riadh* a la même structure que l'EN *stade international du roi Abdelaziz à Riadh*. Ainsi, le mot déclencheur *stade* est remplacé par *hôpital*, le mot déclencheur *roi* est remplacé par *Docteur* et la Tête_EN *Abdelaziz* est remplacée par *Mohamed Hosni*. Aussi, «Fin_EN» garde la même valeur,

«Tête_EN» qui est égale à *international* est remplacée par le mot vide (ε) et le reste de l'EN ne change pas de valeur.

## 3.6. Indépendance du modèle vis-à-vis de la langue

Le modèle proposé est indépendant de la langue utilisée puisque les traits qui le composent ne sont pas influencés par les spécificités des langues.

Nous avons pris le même exemple d'EN décrit dans la **Figure 22** mais traduit respectivement vers le français et l'anglais. Par la suite, nous avons représenté cette EN traduite dans le modèle que nous avons proposé. Nous avons alors obtenu le résultat donné respectivement dans les figures 24 et 25.

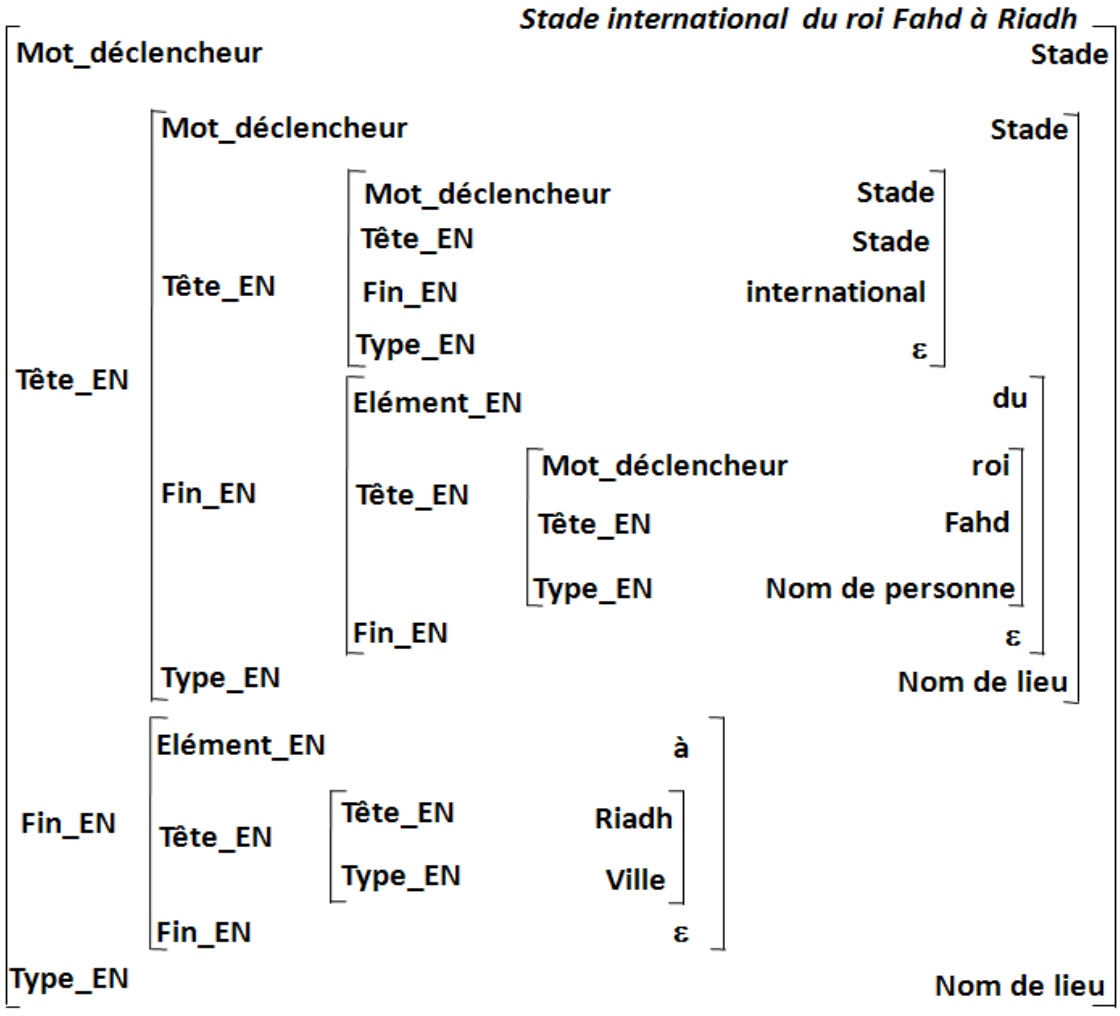


**Figure 24.** *Représentation d'une EN à la langue française*

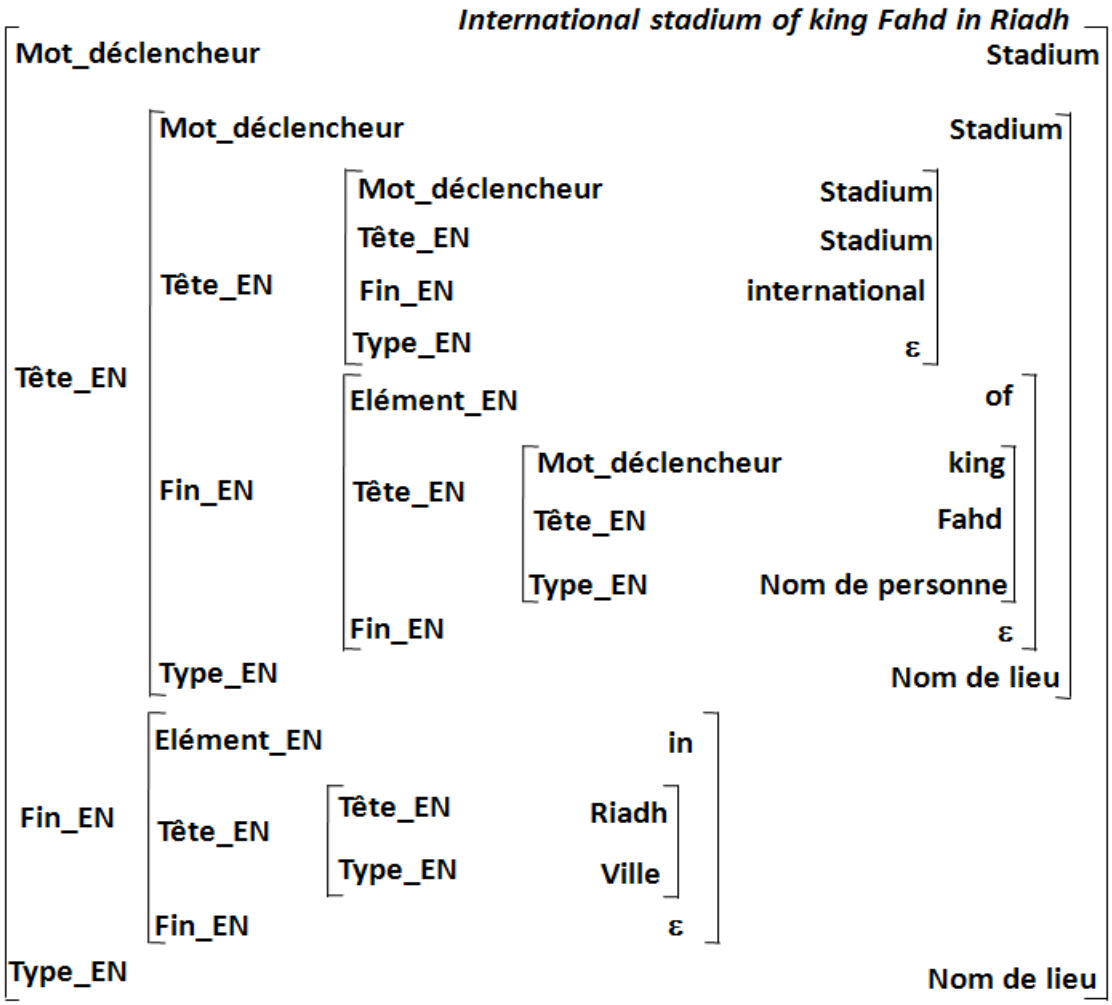


**Figure 25.** *Représentation d'une EN appartenant à la langue anglaise*

Comme illustrées dans les figures 24 et 25, la représentation d'une EN est possible en utilisant le modèle proposé en dépit des variations syntaxique que celle-ci peut subir quand elle est traduite d'une langue à une autre. L'essentiel est que le principe de saturation soit satisfait. Ceci montre aussi l'indépendance syntaxique du modèle.

Les valeurs d'une représentation d'une EN arabe dans le modèle proposé peuvent être traduites vers une langue cible ; dans notre cas le français. La traduction peut être effectuée avec des dictionnaires bilingues sans aucun risque de perte d'informations. Il est évident que la catégorie de l'EN est la même quelque soit la langue.

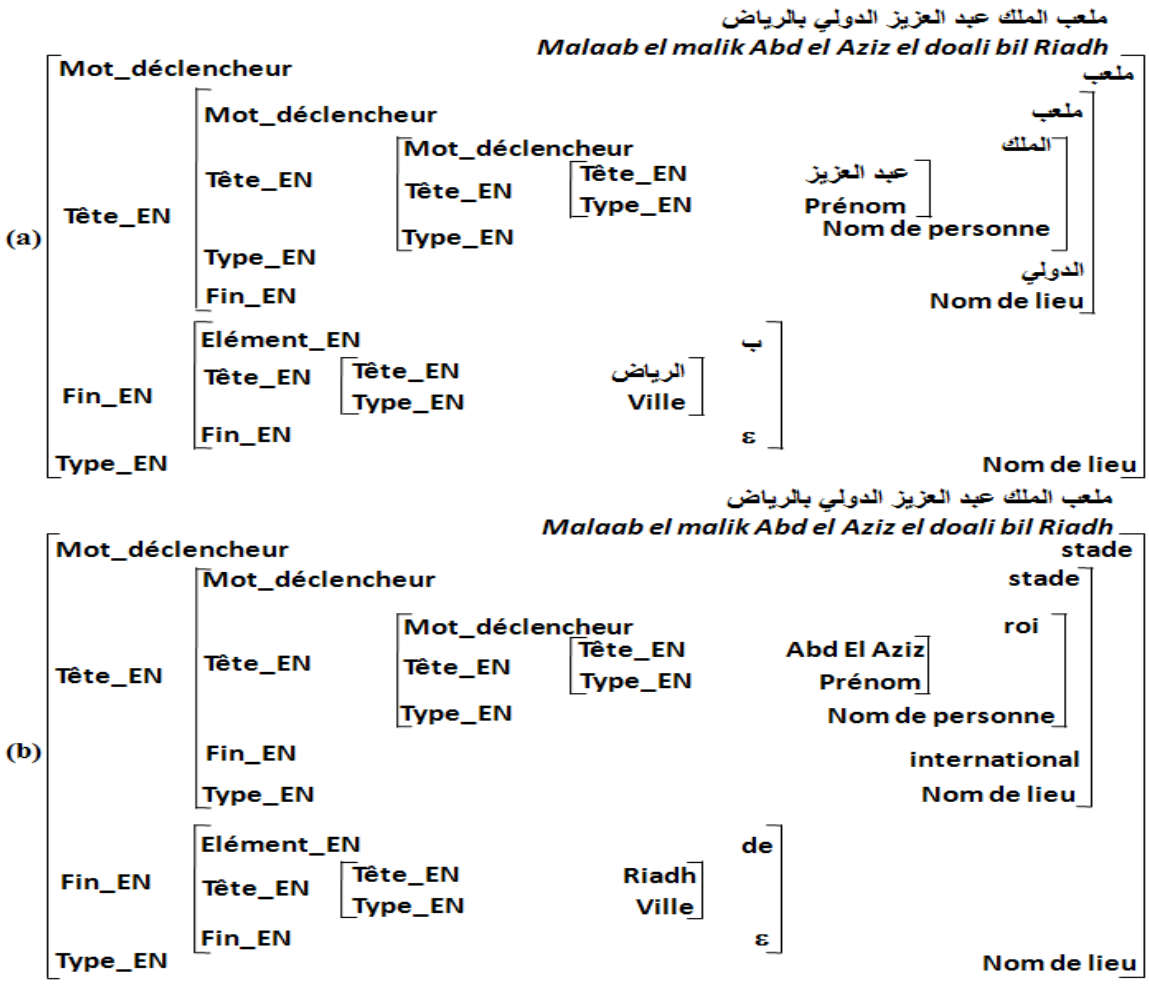


**Figure 26.** *Exemple de traduction mot à mot*

Comme l'indique la **Figure 26** ((a) et (b)), le mot ملعب *malaab* est traduit par *stade*, le mot الملك *almalik* par *roi*, l'adjectif الدولي *elduwaly* par *international et la préposition* ب *bi* par *de*. Il est évident que les types gardent les mêmes valeurs quelque soit la langue. Remarquons bien que la représentation de la traduction mot à mot n'est pas suffisante pour engendrer une EN bien formée dans la langue cible. En effet, la traduction de l'EN de la **Figure 26** (a) ملعب الملك عبد العزيزالدولي بالرياض *mal`ab almalik `abd al`aZyZ elduwaly bilriyaaD* donne dans (b), la succession des mots *stade du roi Abd el Aziz international à Riadh* représentant une EN mal formée car elle ne respecte pas les spécificités de la langue cible. Par conséquent, des règles de réorganisation et réajustement sont nécessaires. Ces règles sont spécifiques à chaque langue. Par exemple, pour le français, il faut que l'adjectif suit le nom avec lui il s'accorde. Par contre, en anglais, il faut que l'adjectif soit avant le nom qu'il décrit.

# Conclusion

Dans ce chapitre, nous avons commencé par la présentation des sources d'inspiration du modèle de représentation formelle que nous avons proposé. Ensuite, nous avons décrit sa structure, l'ensemble de traits qu'il contient ainsi que les principes qu'il doit satisfaire. Cette description nous a permis d'avoir une idée sur la représentation d'une EN dans le modèle proposé. Puis, nous avons vu la possibilité de représenter des EN d'autres langues telles que le français et l'anglais dans le même modèle. Ces représentations nous ont permis de déduire que l'ensemble de traits suggérés sont suffisants pour décrire n'importe quelle EN indépendamment de la catégorie, du domaine et de la langue.

Le modèle formel de représentation des EN arabes peut aider à identifier les dictionnaires et les grammaires nécessaires pour la reconnaissance des EN comme nous le détaillons dans le chapitre suivant.

# Chapitre 4 : Démarche proposée pour la reconnaissance des EN arabes

Ce chapitre est dédié à la présentation de la démarche suivie pour la reconnaissance des EN arabes du domaine du sport. Cette démarche est élaborée en se basant sur le modèle formel de représentation des EN arabes. Ce formalisme permet l'identification des ressources nécessaires pour la reconnaissance (lexiques et grammaires). Les grammaires sont décrites sous forme de transducteurs à états finis facilement implémentés dans l'environnement de développement linguistique NooJ (Silbeztein, 2004) que nous avons utilisé pour valider et expérimenter l'approche proposée. Les différents lexiques sont instanciés à partir du corpus.

Dans le présent chapitre, nous commençons par une présentation générale de la démarche proposée pour la reconnaissance des EN arabes. Ensuite, nous détaillons l'identification et la construction des dictionnaires nécessaires pour la reconnaissance des noms de lieux sportifs. Enfin, nous décrivons les transducteurs identifiés ainsi que la méthode adoptée pour leur construction.

## 1. Présentation générale de la démarche

La méthode de reconnaissance des EN arabes que nous préconisons est basée sur des règles. Ces règles qui sont construites manuellement expriment la structure de l'information à reconnaître et prennent la forme de transducteurs qui seront par la suite directement implémentés dans la plateforme linguistique NooJ (Silbeztein, 2004). Ces transducteurs exploitent généralement des informations d'ordre morphosyntaxique, ainsi que celles contenues dans des ressources (lexiques ou dictionnaires). De plus, ils permettent la description d'enchaînements possibles des constituants des EN arabes appartenant au domaine du sport et particulièrement à la catégorie Nom de lieu. Il est à noter que la reconnaissance des noms de lieux sportifs nécessite la reconnaissance d'autres types d'EN imbriquées que ce soit du même domaine (noms de joueurs, noms d'équipes, noms d'arbitres, noms de sports, etc.) ou d'un autre domaine (noms de personnalité, noms de ville, entités numériques, etc.).

Les étapes de la démarche proposée sont illustrées dans la **Figure 27**.

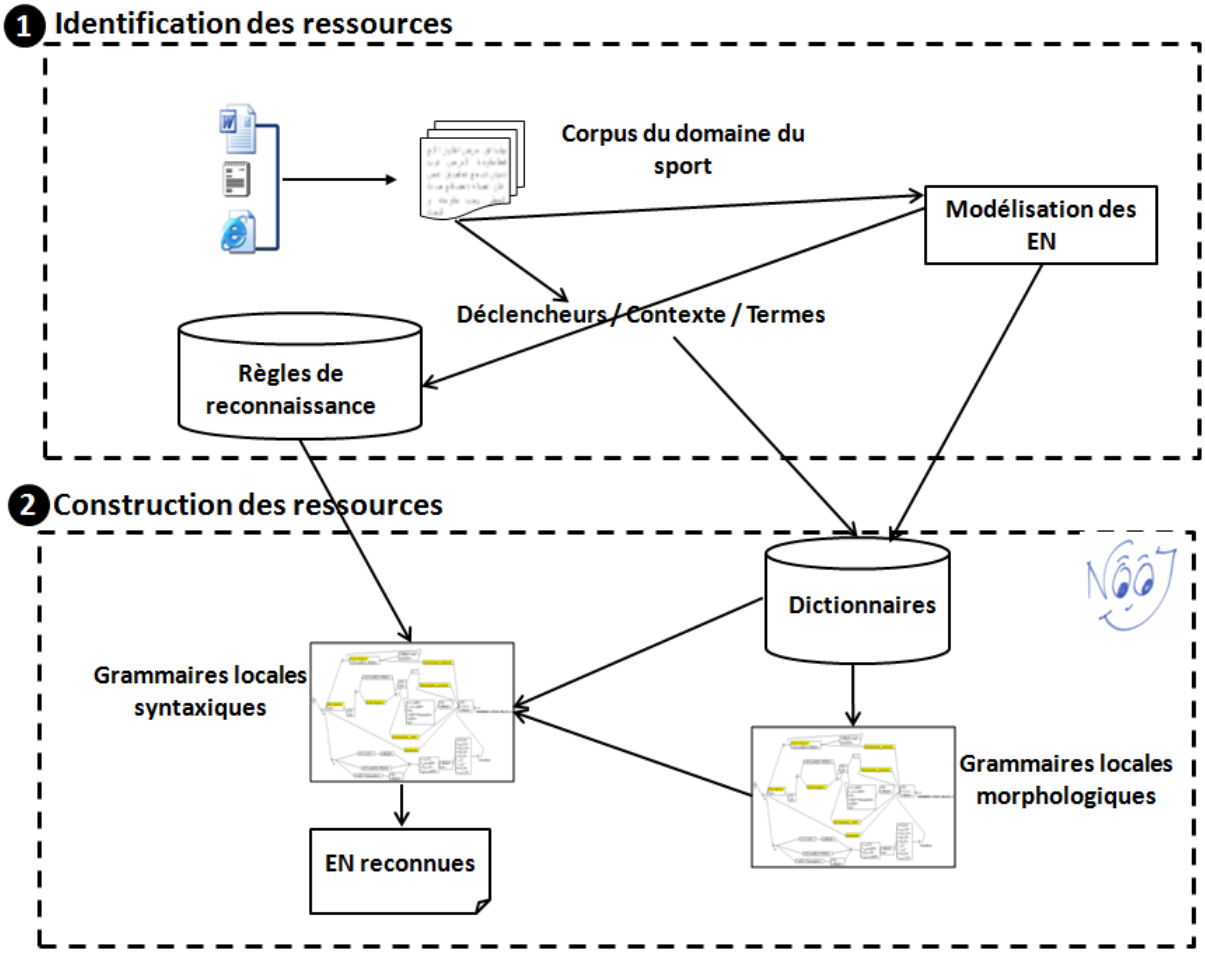


**Figure 27.** *Phases de reconnaissance des EN*

Comme l'indique la **Figure 27**, la démarche de reconnaissance est composée des étapes d'identification des dictionnaires et des grammaires (ou patrons syntaxiques) à partir du modèle de représentation des EN arabes et la transformation des grammaires en transducteurs.

# 2. Identification des dictionnaires

Le modèle proposé a aidé à identifier les dictionnaires nécessaires pour la reconnaissance des EN arabes et ce en associant à chaque EN élémentaire (i.e., valeur du trait «Tête_EN» est élémentaire) un dictionnaire (ex., nom de ville).

Pour le domaine du sport, objet de notre étude, nous avons recensé les dictionnaires suivants :

- Un dictionnaire pour les noms simples
- Un dictionnaire pour les adjectifs (ex., أولمبي *'uwlamby*, وطني *waTany*)

- Un dictionnaire pour les noms d'équipes (ex., الملعب التونسي *almal`ab eltuwnisy*)
- Un dictionnaire pour les noms de joueurs
- Un dictionnaire pour les noms de sport (ex., كرة قدم *korat qadam*)
- Un dictionnaire pour les toponymes
- Un dictionnaire pour les jours et les mois
- Un dictionnaire pour les prénoms
- Un dictionnaire pour les noms de personnalités (ex., خالد بن الوليد *khaalid bin alwalyd*)
- Un dictionnaire pour les mots déclencheurs du domaine (ex., ملعب *mal`ab*)
- Un dictionnaire pour les fonctions (ex., أمير *'amyr*)

La structure des entrées des différents dictionnaires n'est pas la même. Elle peut varier d'un dictionnaire à un autre mais contient au minimum :

- la catégorie grammaticale de l'entrée (Nom, Adjectif) et
- le trait sémantique qui définit le type de l'entrée (Fonction, Prénom, nom équipe, nom joueur,….)

A ces informations on trouve selon le dictionnaire des informations additionnelles comme :

- le genre (féminin ou masculin) et le nombre (singulier, duel et pluriel)
- le modèle de dérivation pour reconnaitre les formes dérivées du lemme contenue dans l'entrée
- le modèle flexionnel pour reconnaitre les formes fléchies du lemme contenue dans l'entrée.
- Le trait détermination pour les noms qui acceptent d'être déterminés tels que الرياض *elriyaD*. Ce n'est pas le cas par exemple du toponyme تونس tuwnis.

Toutes ces informations sont exploitées par les transducteurs pour guider le processus de reconnaissance et lever certaines ambiguïtés.

Notons que, pour effectuer la traduction des EN arabes, nous avons ajouté pour les entrées de tous les dictionnaires (à l'exception de celui des prénoms) un trait qui représente la traduction en Français du lemme.

# 3. Identification et construction des transducteurs

Le modèle formel, que nous avons proposé, a aidé aussi à identifier les transducteurs nécessaires pour la reconnaissance des EN. En effet, chaque trait «Tête_EN», dont la valeur est structurée et composée par des traits non vides, autre que le trait «Type_EN», sera transformé en une grammaire syntaxique (nom de lieu, nom de personne, ...). La construction de ces grammaires est réalisée à l'aide des patrons syntaxiques pour faciliter la transformation en transducteurs.

Dans ce qui suit, nous détaillons les parons syntaxiques déduits du modèle de représentation et ensuite nous donnons les transducteurs correspondants.

## 3.1. Identification des patrons syntaxiques

Pour faciliter l'identification des transducteurs nécessaires pour la reconnaissance des EN, nous avons transformé les différentes structures attribut-valeur du modèle de représentation des EN arabes, en patrons syntaxiques. En effet, les patrons syntaxiques donnent l'agencement des différents constituants des EN d'une manière linéaire facilement transposable sous forme de graphes. Alors que la structure attribut-valeur ne possède pas cette propriété.

Du modèle de représentation, nous distinguons sept patrons syntaxiques qui décrivent les EN arabes de type nom de lieu sportif. Ces patrons sont dégagés selon les constituants communs d'une EN arabe. Dans ce qui suit, nous les détaillons.

Le patron 1 décrit les différentes formes d'un nom de personnalité. Ce patron peut être appelé par d'autres patrons qui concernent les noms de lieux sportifs.

| <Patron 1> := [<Fonction>] <Prénom> [Fils de <Prénom>]* [<Nom>] |
|---|

Parmi les formes générées par Patron 1, nous pouvons citer les exemples suivants :

(1) ملعب **الطيب المهيري** *mala`ab* ***elTayib almhyry*** <Nom> <Prénom>

(2) ملعب **العقيد لطفي** *mal`ab* ***al`aqyd lotfii*** <Fonction> <Prénom>

(3) إستاد مدينة **الأمير سلطان بن عبد العزيز** الرياضية *'istaad mdynat* ***al'amyr Soltaan bin `abd al`aZiZ*** *elriyDiyyah* <Fonction> <Prénom> Fils de <Prénom>

(4) مسبح **الأسد** الدولي *masbah* ***al'asad*** *elduwaly* <Prénom>

(5) حمص – **خالد بن الوليد** ملعب *mal`ab* ***khaalid bin alwalyd*** – homos <Prénom> Fils de <Prénom>

(6) **الأمير عبد الله الفيصل** ملعب *mal`ab al'amyr* ***abd allah al-faysal*** <Fonction> <Nom> <Prénom>

(7) **الأمير محمد بن عبد العزيز** ملعب *mal`ab* ***al'amyr muHammad bin `abd al`aZiZ*** <Fonction> <Prénom> fils de <Prénom>

Le patron 2 décrit les EN où le nom de personnalité constitue un élément obligatoire. Il représente les EN qui commencent par un ou plusieurs mots déclencheurs suivi par un certain nombre d'adjectifs suivi par un nom de personnalité suivi ou non par un toponyme. Il décrit aussi les EN qui commencent par un ou plusieurs mots déclencheurs suivi d'un nom de personnalité suivi ou non par des adjectifs suivi ou non par un ou plusieurs toponymes.

<Patron 2> := <Mot_déclencheur>$^{+}$ ([<Adjectif>]$^{*}$ <Patron 1> | <Patron 1> [<Adjectif>]$^{*}$) [<Toponyme>]$^{*}$

Parmi les EN, basées sur l'existence d'un nom de personnalité et générées par Patron 2, nous citons les suivantes :

(8) ملعب مبارك بوصيف *mal`ab mubaarak bouSyf* : cet exemple est représenté par : <Mot_déclencheur> <Patron 1> où le mot déclencheur est ملعب *mal`ab* et le nom de personnalité est مبارك بوصيف *mubaarak bouSif* .

(9) ملعب الطيب المهيري بصفاقس *mal`ab elTayib almhyry biSafaaqus*: cet exemple est sous la forme <Mot_déclencheur> <Patron 1> <Toponyme> où le mot déclencheur est ملعب *malaab*, le nom de personnalité est الطيب المهيري *elTayib almhyry* et le toponyme est صفاقس *Safaaqus*.

(10) المركب الرياضي ببرج السدرية *almurakkab elriyaaDy biborj elsidriyya*: cet exemple respecte la règle suivante : <Mot_déclencheur> <Adjectif> <Toponyme> où le mot déclencheur est المركب *almurakkab*, l'adjectif est الرياضي *elriyaaDy* et le toponyme est برج السدرية *borj elsidriyya*.

(11) المركب الرياضي محمد الخامس بالدار البيضاء *almurakkab elriyaaDy muHammad alkhaamis bildaar albayDaae*: Cet exemple respecte la règle suivante : <Mot_déclencheur> <Adjectif> <Patron 1> <Toponyme> où le mot déclencheur est المركب *almurakkab*, l'adjectif est الرياضي *elriyaaDy*, le nom de personnalité est محمد الخامس *muHammad alkhaamis* et le toponyme est الدار البيضاء *eldaar albayDaae*.

(12) مجمع السلطان قابوس الرياضي ببوبش *majma` elSolTaan qaabuws elriyaDy bibuwbish*: cet exemple respecte la règle suivante : <Mot_déclencheur> <Patron 1> <Adjectif> <Toponyme> où le mot déclencheur est مجمع *majma`*, le nom de personnalité est السلطان قابوس *elSolTaan qaabuws*, l'adjectif est الرياضي *elriyaDy* et le toponyme est بوبش *buwbish*.

(13) إستاد مدينة الأمير سلطان بن عبد العزيز الرياضية (المحالة) *'istaad madynat al'amyr soltaan bin `abd al`aZiZ elriyaaDiyya (almouHalah)*: cet exemple s'écrit sous la forme de <Mot_déclencheur> <Mot_déclencheur> <Patron 1> <Adjectif> <Toponyme>. Dans cet exemple, les mots déclencheurs sont إستاد *'istaad* et مدينة *madynat*, le nom de personnalité est الأمير سلطان بن عبد العزيز *al'amyr soltaan bin `abd al`aZiZ*, l'adjectif est الرياضية *elriyaaDiyya* et le toponyme est المحالة *almouHalah*.

Le patron 3 décrit alors toutes les EN qui commencent par un mot déclencheur suivi ou non par des adjectifs suivi par un ou plusieurs toponymes suivi ou non par un nom de sport. Il décrit aussi toutes les EN qui commencent par un mot déclencheur suivi par un toponyme suivi par un ou plusieurs adjectifs suivi ou non par un nom de sport.

| <Patron 3> := <Mot_déclencheur> ([Adjectif]$^{*}$ <Toponyme>$^{+}$ \| <Toponyme> [<Adjectif>$^{+}$]) [<NomSport>] |
|---|

Du Patron 3, nous distinguons les formes générées suivantes :

(14) إستاد لاسي *'istaad laasy*: cet exemple respecte la forme suivante : <Mot_déclencheur> <Toponyme> où le mot déclencheur est إستاد *'istaad* et le toponyme est لاسي *laasy*.

(15) ملعب بانياس – بانياس *mal`ab baaaniyaas - baaniyaas*: cet exemple est sous la forme <Mot_déclencheur> <Toponyme> <Toponyme> où le mot déclencheur est ملعب *mal`ab*, le premier toponyme est بانياس *baaaniyaas* et le deuxième toponyme est aussi بانياس *baaaniyaas*.

(16) الملعب الاولمبي بسوسة *almal`ab al'uwlimpy bisuwsa*: cet exemple est représenté par : <Mot_déclencheur> <Adjectif> <Toponyme> où le mot déclencheur est الملعب *almal`ab*, l'adjectif est الاولمبي *al'uwlimpy* et le toponyme est سوسة *suwsa*.

(17) إستاد البحرين الوطني *'istaad albaHrayn alwaTany*: cet exemple s'écrit sous la forme : <Mot_déclencheur> <Toponyme> <Adjectif> où le mot déclencheur est إستاد *'istaad*, le toponyme est البحرين *albaHrayn* et l'adjectif est الوطني *alwaTany*.

(18) ملعب طرابلس الاولمبي الدولي *mal`ab Taraabols al'uwlimby elduwaly*: cet exemple respecte la règle suivante : <Mot_déclencheur> <Toponyme> <Adjectif> <Adjectif>

où le mot déclencheur est ملعب *mal`ab*, le toponyme est طرابلس *Taraabols* et les deux adjectifs sont respectivement الاولمبي *al'uwlimby* et الدولي *elduwaly*.

(19) المركب الرياضي البلدي لقصبة تادلة *almurakkab elriyaaDy albalady liqaSabat taadilah*: cet exemple est sous la forme : <Mot_déclencheur> <Adjectif> <Adjectif> <Toponyme> où le mot déclencheur est المركب *almurakkab*, les adjectifs sont respectivement sont الرياضي *elriyaaDy* et البلدي *albalady* et le toponyme est قصبة تادلة *qaSabat taadilah*.

(20) ملعب دبي للتنس *mal`ab dubay litinnis*: cet exemple à la forme : <Mot_déclencheur> <Toponyme> <NomSport>. Dans cet exemple, le mot déclencheur est ملعب *mal`ab*, le toponyme est دبي *dubay* et le nom de sport est التنس *eltinnis*.

Le patron 4 représente les EN qui commencent par un mot déclencheur suivi par un nom d'équipe suivi par un nom de joueur ou un nom de personnalité. Dans ce patron le nom d'équipe est le constituant obligatoire.

| <Patron 4> := <Mot_déclencheur> <Nom_équipe> ( [<Nom_joueur>] \| [<Patron 1>]) |
|---|

Parmi les EN générées par Patron 4, nous trouvons les suivantes :

(21) ملعب نادي القادسية *mal`ab naadii alqaadisiyyah* : cet exemple est représenté sous cette forme : <Mot_déclencheur> <Nom_équipe> où le mot déclencheur est ملعب *mal`ab* et le nom d'équipe est نادي القادسية *naadii alqaadisiyyah*.

(22) ستاد الزمالك (حلمي زامورا) *staade elZamaalik (Hilmy Zaamouraa)*: cet exemple est sous la forme : <Mot_déclencheur> <Nom_équipe> <Nom_joueur> où le mot déclencheur est ستاد *staade*, le nom d'équipe est الزمالك e*lZamaalik* et le nom de joueur est حلمي زامورا *Hilmy Zaamouraa*.

(23) ستاد المقاولون العرب – عثمان احمد عثمان *staade almuqqawiloun al`arab – `uthmaan 'ahmad `uthmaan*: cet exemple est représenté par : <Mot_déclencheur> <Nom_équipe> <Patron 1> où le mot déclencheur est ستاد *staade*, le nom d'équipe est المقاولون العرب *almuqqawiloun al`arab* et le nom de personnalité est عثمان احمد عثمان *`uthmaan 'ahmad `uthmaan*.

Il arrive des cas où l'EN doit contenir un nom de personnalité qui est placé avant le nom d'équipe telles que صالة عبد العزيز بنادي العربي الخطيب *Saalat `abd al`aZyZ alkhatyb binaady al`araby* où le mot déclencheur est صالة *Saalat*, le nom de personnalité est عبد العزيز الخطيب

*`abd al`aZyZ alkhatyb* et le nom d'équipe est نادي العربي *naady al`araby*. A ce moment l'ajout d'un autre patron s'avère nécessaire. Ce patron est le suivant :

| <Patron 5> := <Mot_déclencheur> <Patron 1> <Nom_équipe> |
|---|

Le patron 5 représente alors les EN qui commencent par un mot déclencheur suivi par un nom de personnalité suivi par un nom d'équipe.

Le patron 6 représente les EN qui commencent par un mot déclencheur suivi d'un nom commun suivi ou non par un adjectif suivi ou non par un toponyme.

| <Patron 6> := <Mot_déclencheur> <Nom_commun> [<Adjectif>] [<Toponyme>] |
|---|

Parmi les EN générées par Patron 6, nous citons les suivantes :

(24) ملعب التوحيد *mal`ab eltawHyd*: cet exemple est sous la forme <Mot_déclencheur> <Nom_commun> où le mot déclencheur est ملعب *mal`ab* et le nom commun est التوحيد *eltawHyd*.

(25) إستاد الجيش الملكي *'istaad aljaysh almalaky*: cet exemple est sous la forme de <Mot_déclencheur> <Nom_commun> <Adjectif> où le mot déclencheur est إستاد *'istaad*, le nom commun est الجيش *aljaysh* et l'adjectif est الملكي *almalaky*.

(26) صالة الاتحاد (بالدعية) *Saalat al'itiHaad (bilda`ya)*: cet exemple est sous la forme de <Mot_déclencheur> <Nom_commun> <Toponyme> où le mot déclencheur est صالة *Saalat*, le nom commun est الاتحاد *al'itiHaad* et le toponyme est الدعية *elda`ya*.

Le patron 7 représente les EN qui commencent par un mot déclencheur suivi d'une date suivi ou non d'un toponyme. Une date dans ce patron est composée du jour suivi du mois suivi ou non de l'année suivi ou non par un toponyme.

| <Patron 7> := <Mot_déclencheur> <Jour> <Mois> [<Année>] [<Toponyme>] |
|---|

Notons que la forme de la date représentée est uniquement celle qui peut exister dans une EN de type nom de lieu. Comme exemples générées du Patron 7, nous trouvons les suivants :

(27) قاعة 3 مارس *qaa`at 3 maaaris*: cet exemple set sous la forme <Mot_déclencheur> <Jour> <Mois> où le mot déclencheur est قاعة *qaa`at*, le jour est 3 et le mois est مارس *maaaris*.

(28) ملعب 7 نوفمبر برادس *mal`ab 7 nuwfambar biraadis*: cet exemple est représenté par la règle suivante : <Mot_déclencheur> <Jour> <Mois> <Toponyme> où le mot déclencheur est ملعب *mal`ab*, le jour est 7, le mois est نوفمبر *nuwfambar* et le toponyme est رادس *raadis*.

(29) 1956 ماي 19 ملعب *mal`ab 19 maay 1956*: cet exemple est sous la forme de <Mot_déclencheur> <Jour> <Mois> <Année> où le mot déclencheur est ملعب *mal`ab*, le jour est 19, le mois est ماي *maay* et l'année est 1956.

(30) باتنة -1954 نوفمبر 7 ملعب *mal`ab 7 nuwfambar 1954 - baatinah*: cet exemple est sous la forme <Mot_déclencheur> <Jour> <Mois> <Année> <Toponyme> où le mot déclencheur est ملعب *mal`ab*, le jour est 7, l'année est 1954 et le toponyme est باتنة *baatinah*.

Les sept patrons énumérés auparavant seront formalisées ultérieurement par des grammaires syntaxiques à l'aide des transducteurs.

## 3.2. Transformation des patrons syntaxiques en transducteurs

Cette étape consiste à formaliser les règles déjà construites dans l'étape de l'étude du corpus en utilisant le formalisme des transducteurs. Chaque transducteur est caractérisé par un nœud initial et un nœud final . Nous effectuons la lecture de ces transducteurs de droite à gauche vu qu'il s'agit de la langue arabe. La **Figure 28** représente le transducteur principal pour la reconnaissance des noms de lieux sportifs. On distingue au total 5 graphes qui représentent les différents types de lieux sportifs (Stade, Complexe, piscine, salle et cité) conformément à la hiérarchie des EN proposée dans le chapitre 2.

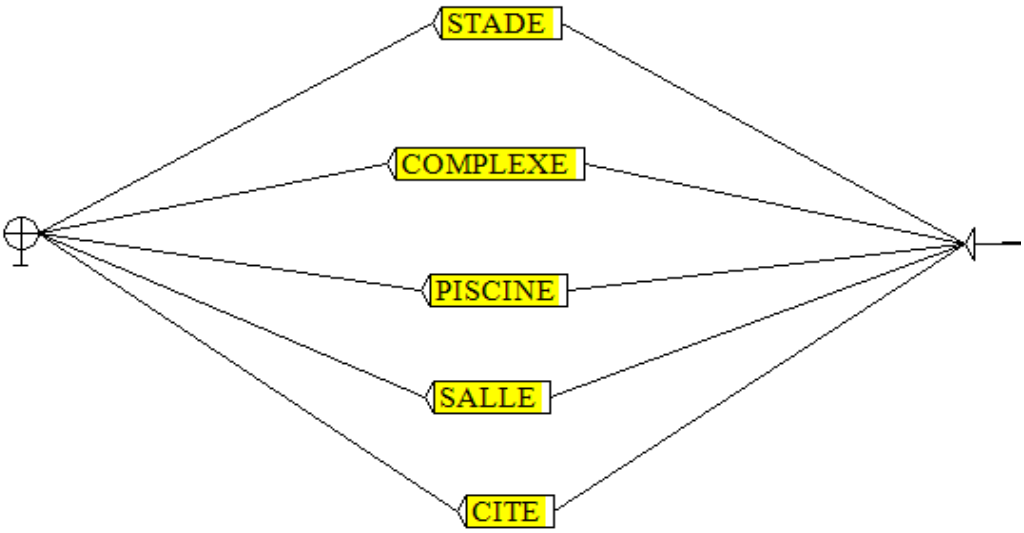


**Figure 28.** *Transducteur principal pour la reconnaissance de EN du domaine du sportif*

Notons que le transducteur de la **Figure 28** contient au total 26 sous-graphes. Ces sous-graphes représentent les EN imbriquées contenues dans l'EN principale et données par le

modèle de représentation formelle des EN proposé dans le chapitre précédent. Chaque chemin d'un sous-graphe décrit un patron syntaxique.

Pour illustrer cette transformation, nous détaillons dans ce qui suit les sous graphes qui sont imbriqués dans le nœud STADE qui représente celui le plus riche de tous les nœuds.

### 3.2.1. Reconnaissance des noms de stade

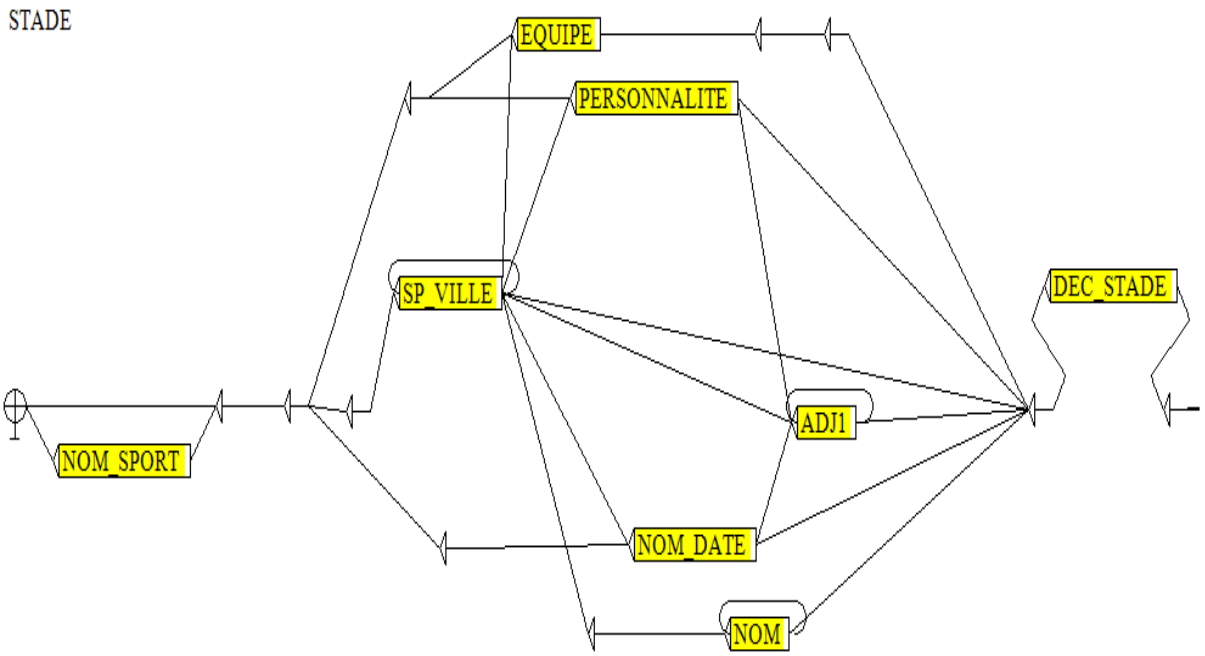


**Figure 29.** *Transducteur de reconnaissance des EN de la catégorie Stade*

Le transducteur de la **Figure 29** montre qu'un nom de stade peut contenir un nom d'une personnalité, un nom, un adjectif, une ville, une catégorie géographique ou une date. Un nom d'une personnalité peut être précédé par un adjectif ou non. Ce qui exprime le nombre d'entrées qui est égal à deux au nœud « PERSONNALITE». Après le nom de personnalité, le nom d'un stade peut être suivi par un nom d'une ville (présenté par le sous-graphe « SP-VILLE »), un nom d'équipe (représenté par le sous-graphe « EQUIPE ») ou un nom de sport (représenté par le sous-graphe NOM_SPORT) ou rien (du nœud PERSONNALITE, on arrive directement à un état final). La combinaison de chaque entrée avec chaque sortie modélise une règle. Ceci montre que pour reconnaître les EN de type *Nom de lieu*, nous avons construit d'autres graphes permettant la reconnaissance d'autres types d'EN. Dans ce qui suit, nous donnons une idée sur les sous-graphes qui permettant la reconnaissance d'autre type d'EN et qui sont utiles pour la reconnaissance des noms de lieux sportifs.

### 3.2.2. Reconnaissance des noms de personnalité

Le transducteur de la **Figure 30** permet la reconnaissance des noms de personnalités. Ce transducteur décrit les différentes possibilités pour la formation d'un nom de personnalité (Patron 1).

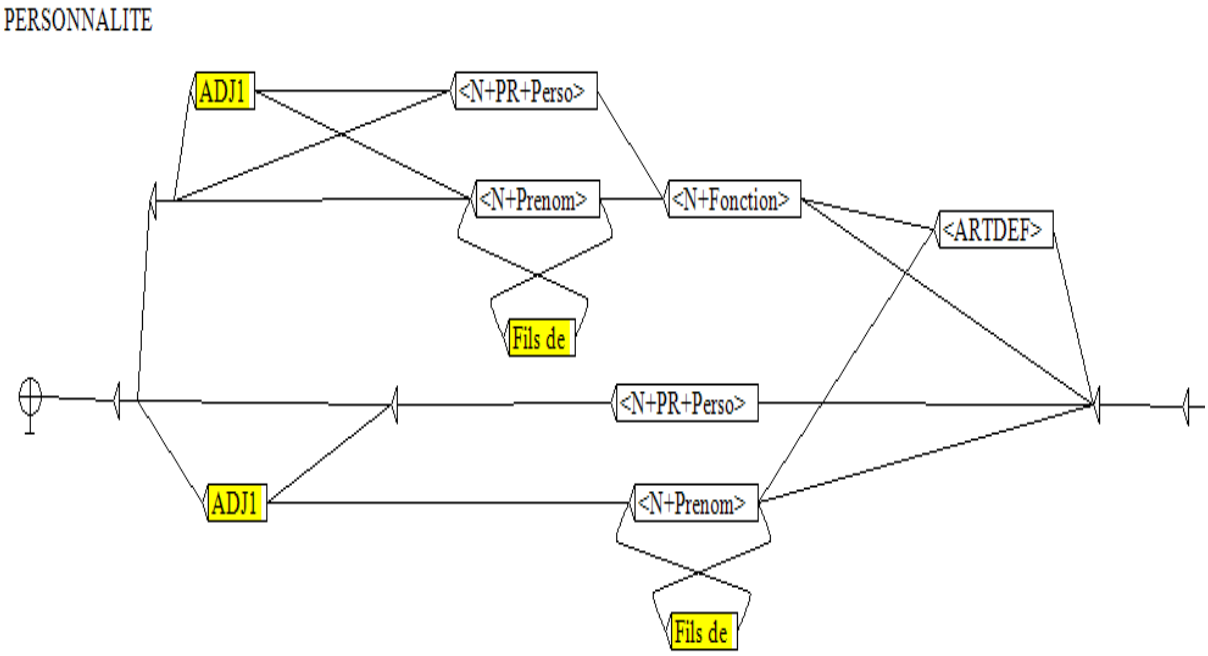


**Figure 30.** *Transducteur de reconnaissance des noms de personnalité*

Comme le montre le transducteur de la **Figure 30**, un nom de personnalité peut exister sous différentes formes. En effet, il peut être composé uniquement par un prénom (N+Prenom) ou par un nom et prénom (représenté par le trait Perso) comme il peut être précédé par un nom de fonction (N+Fonction) tel que الأمير *al'amyr le prince*. Le nœud contenant «ARTDEF» indique l'article défini « ال » qui peut être agglutiné au nom de fonction ou au prénom tel que الباسل *albaasil*.

### 3.2.3. Reconnaissance des toponymes

La plupart des EN arabes du domaine du sport ayant pour type *nom de lieu* contiennent un toponyme. Ce toponyme est précédé généralement par un mot déclencheur. Le transducteur de la **Figure 31** décrit les différentes possibilités pour la formation d'un toponyme.

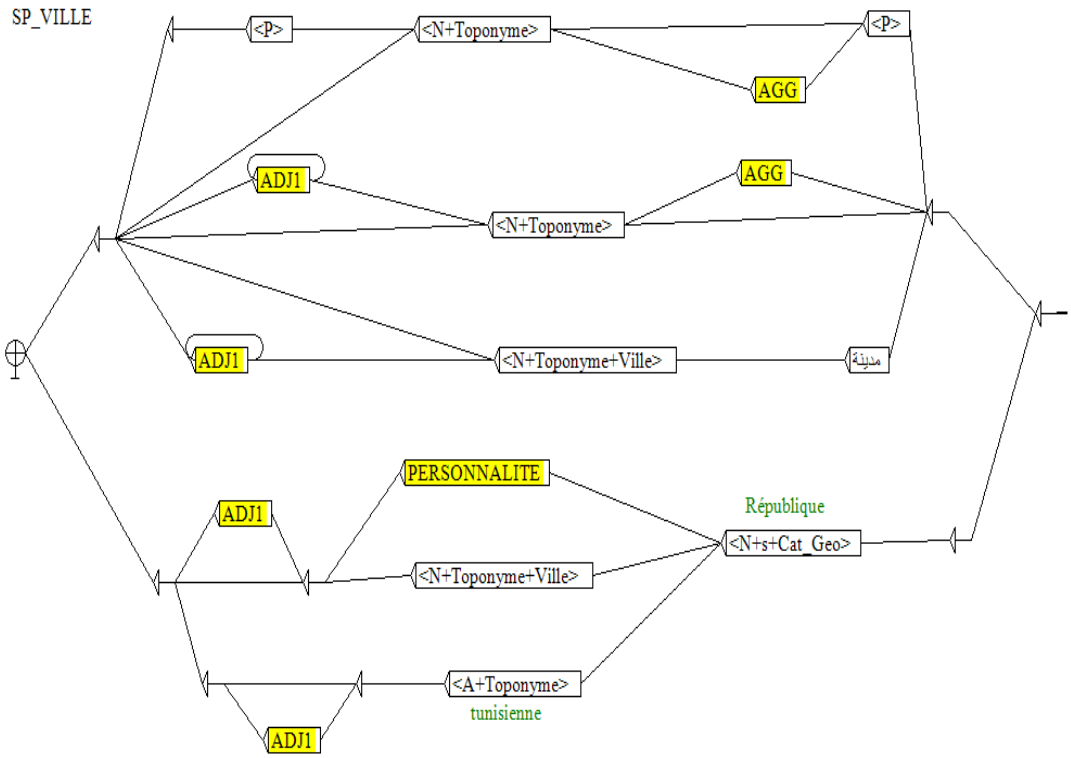


**Figure 31.** *Transducteur de reconnaissance des toponymes*

Comme le montre le transducteur de la **Figure 31**, un toponyme peut être précédé par un caractère spécial tel que) et -. Ce caractère est décrit par <P>. Il peut être précédé aussi par un mot déclencheur. Si le mot déclencheur est égal à مدينة *madynat ville* alors le toponyme doit être une ville. Ceci est représenté par le nœud N+Toponyme+Ville. Le mot déclencheur peut être aussi une catégorie géographique telle que محافظة *muHaafaDa* et جمهورية *jomhuwriyya*.

Egalement, le toponyme peut être composé par un nom d'une personnalité tel que مدينة الملك عبد العزيز الرياضية *madynat almalik `abd al`aZyZ elriyaaDiyya* (*cité sportive du roi Abdelaziz)* et suivi par un adjectif.

### 3.2.4. Reconnaissance des dates

Les noms de lieux en général peuvent contenir des dates. Ces dates font référence à des événements qui ont eu lieu dans le passé. Par exemple dans l'EN ملعب 15 أكتوبر *stade 15 octobre*, la date fait référence à la fête de l'évacuation (Aïd El Jala). Le transducteur de la **Figure 32** décrit les différentes formes sous lesquelles une date peut exister dans une EN de type nom de lieu et en particulier du domaine du sport.

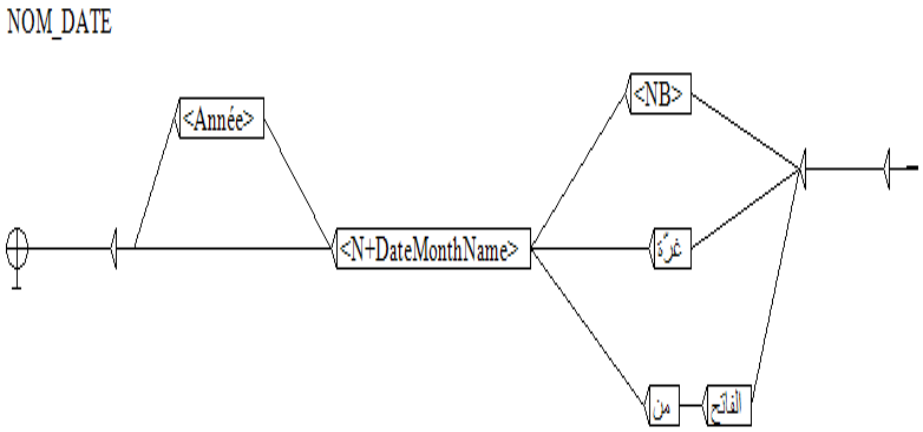


**Figure 32.** *Transducteur de reconnaissance des dates*

Le transducteur de la **Figure 32** traite uniquement les formes de dates qui peuvent exister dans les noms de lieu liés au domaine du sport. Ces dates sont sous cette forme : jour suivi du mois suivi ou non de l'année. En effet, on ne peut pas trouver par exemple un nom de lieu qui contient une date sous cette forme : jour suivi de / suivi du mois suivi de / suivi de l'année. Cependant ce transducteur peut être enrichi et réutilisé selon l'objectif à atteindre.

# Conclusion

Dans ce chapitre, nous avons présenté la démarche que nous avons suivie pour la reconnaissance des EN arabes et particulièrement les noms de lieux sportifs. Cette démarche, qui se base sur le modèle de représentation des EN arabes, est composé de deux étapes principales : Identification des dictionnaires et des grammaires et la transformation des grammaires identifiées, (exprimées en Patrons syntaxiques) en transducteurs. Le processus de reconnaissance à travers ces transducteurs est guidé par les informations disponibles dans les dictionnaires. Nous avons identifié, pour la reconnaissance des noms de lieux du domaine du sport, 11 dictionnaires et 5 graphes de base contenant au total 26 sous-graphes. Ces différentes ressources sont exploitées aussi pour la traduction des EN reconnues de l'arabe vers le français, qui sera détaillée dans le chapitre suivant.

# Chapitre 5 : Démarche proposée pour la traduction des EN arabes

Le présent chapitre est dédié à la présentation de la démarche proposée pour la traduction des EN de la langue arabe vers la langue française. Cette démarche proposée contient un ensemble d'étapes qui tient en compte les spécificités de l'arabe et du français. Elle doit être indépendante du domaine et de la catégorie grammaticale. Cependant, la traduction de l'arabe vers le français ne conduit pas toujours aux résultats attendus. Pour cette raison, la translittération d'une partie d'une EN arabe se révèle parfois intéressante et peut compléter le sens d'une EN traduite en français. Ainsi, la démarche proposée contient deux grandes phases : une phase de traduction et une phase de translittération.

Dans ce chapitre, nous commençons tout d'abord par présenter les problèmes liés à la traduction des EN arabes vers le français. Ensuite, nous donnons une idée générale sur la méthode proposée pour la traduction. Puis, nous décrivons les différentes phases du processus de traduction. Enfin, nous détaillons le processus de translittération.

## 1. Problèmes liés à la traduction arabe-français des EN

La traduction des EN de l'arabe vers le français n'est pas une tâche triviale. En effet, plusieurs problèmes doivent être traités. Nous les résumons dans les points suivants :

- **La place des adjectifs dans l'EN :** Elle n'est pas la même pour les deux langues. Par exemple, *Malaab Al-malik Faissal* ***al-oulimpi*** se traduit en français par *le stade* ***olympique*** *du roi Faiçal*. Dans la langue arabe, l'adjectif est à la fin de l'EN par contre dans la langue française l'adjectif suit le nom auquel il s'accorde (en deuxième position).
- **Le trait genre (masculin ou féminin) :** Il n'est pas nécessairement identique dans les deux langues (source et cible). Par exemple, le mot «مسبح» *masbaH* est masculin, alors que sa traduction «piscine» est féminine.

- **L'ambiguïté des noms de villes et de capitales :** Par exemple, le mot «تونس, *Tuwnis*» peut être traduit par République Tunisienne ou Tunis, la capitale. C'est aussi le cas de «الجزائر, Aljazaa'ir» qui peut être traduit par Algérie ou Alger.

Ainsi, la traduction des EN de l'arabe vers le français ne peut pas se faire mot à mot car d'une part, les connecteurs (préposition et conjonctions) ne sont pas les mêmes et d'autre part, l'ordre des mots est différent. En outre, la traduction des EN n'est pas toujours suffisante pour compléter le sens d'une EN surtout dans le cas des prénoms. C'est pourquoi l'intégration d'un module de translittération s'avère importante.

Afin de réussir la translittération d'une partie de l'EN, nous tenons compte des deux points suivants :

- **La résolution du problème de voyellation :** En effet, un corpus peut être voyellé ou non. Par conséquent, la translittération obtenue d'un mot n'est pas toujours correcte. Par exemple, si nous voulons translittérer le prénom محمد Mohamed sans le voyeller, nous obtenons un ensemble de consonnes mHmmd ; ce qui est impossible à prononcer. Toutefois, nous devons voyeller ce nom propre avant de le translittérer مُحَمَّد. Ainsi, sa translittération donne le mot Mohammad.
- **La résolution du problème de flexion pour les prénoms :** En effet, en arabe, un nom propre peut être décliné afin de respecter la fonction grammaticale. Cette déclinaison parvient à la fin du nom propre (d'un mot arabe en général). Par exemple, on peut trouver dans un texte محمدا *MuHammadan*, mais en français, les noms propres ne sont pas déclinés.

Après avoir introduit les problèmes liés à la traduction arabe-français des EN, nous présentons dans la section suivante la démarche de traduction proposée.

# 2. Présentation générale de la démarche de traduction proposée

La démarche proposée pour la traduction des EN arabes tient compte des problèmes cités auparavant. Cette démarche s'articule autour de deux phases : une phase de traduction et une phase de translittération. La **Figure 33** illustre l'enchainement de ces deux phases.

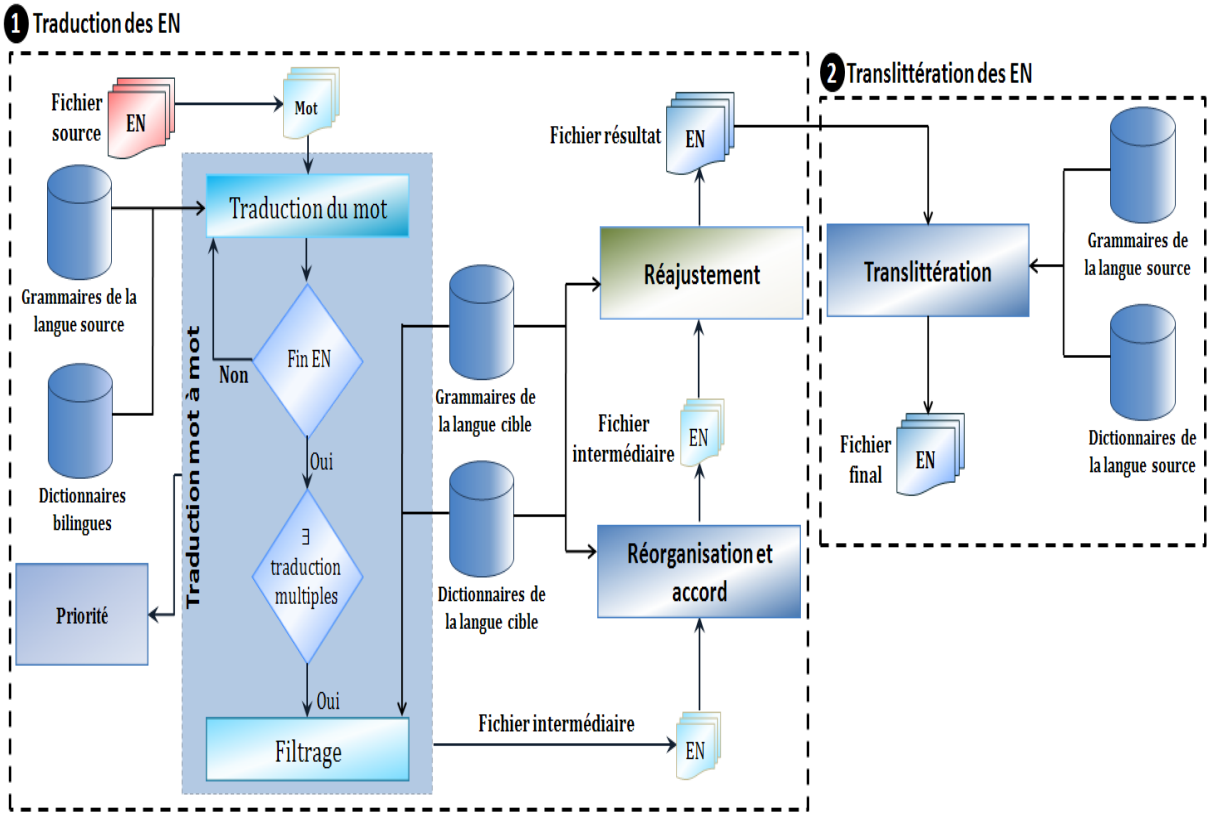


**Figure 33.** *Phases de traduction des EN*

Comme le montre la **Figure 33**, la phase de traduction contient trois étapes : la traduction mot à mot, la réorganisation et accord et le réajustement. Chaque étape utilise comme entrée la sortie de l'étape précédente. L'entrée à la phase de traduction est le fichier contenant les EN déjà reconnues dans le processus de reconnaissance décrit dans le chapitre précédent. Chaque EN de ce fichier va être décomposée et traitée mot par mot. Chaque mot va être traduit en utilisant les grammaires et les dictionnaires bilingues de la langue source. Avant de passer à traduire mot par mot l'EN qui suit, nous vérifions s'il existe des traductions multiples qui ont été attribuées à un même mot. Si c'est le cas, alors nous passons à éliminer ces traductions multiples et ceci en utilisant des dictionnaires et des grammaires de la langue cible. Le même processus s'applique pour le reste des EN. L'étape de traduction mot à mot s'effectue en deux itérations : la première itération pour les mots composés et la deuxième pour les mots simples. La priorité est donnée pour les mots composés car si nous trouvons un mot composé tel que le toponyme حمام الأنف *Hammam al'anf*, et si nous commençons par sa traduction mot par mot alors nous allons obtenir comme résultat « bain de nez » ce qui est faux et change tout le sens de l'EN. La sortie de l'étape de traduction mot à mot est un fichier intermédiaire contenant le même nombre d'EN que le fichier source mais traduites mot à mot sans tenir compte des spécificités de la langue cible. Ce fichier va être l'entrée pour l'étape de réorganisation et

accord. Dans cette étape, nous réorganisons les constituants de chaque EN en déplaçant chaque adjectif juste après le nom auquel il s'accorde et en effectuant en même temps l'accord. Ce traitement est réalisé en utilisant des dictionnaires et des grammaires de la langue cible. La sortie de cette étape est aussi un fichier intermédiaire qui va être une entrée dans l'étape suivante : l'étape de réajustement. Cette étape consiste à ajouter des prépositions entre deux noms successifs ou entre un adjectif et un nom si c'est nécessaire. L'insertion des prépositions s'effectue en utilisant des règles de réajustement qui sont modélisées par des grammaires et des dictionnaires de la langue cible. La sortie de cette étape est un fichier contenant toujours le même nombre d'EN mais bien traduites en tenant compte des spécificités de la langue française. Dans ces EN, il y a des mots qui ont resté dans la langue source tel que les prénoms. Ainsi, le fichier obtenu va être une entrée pour la phase de translittération afin de translittérer les mots qui n'ont pas été traduits. Cette phase se sert des grammaires et des dictionnaires de la langue source.

Dans ce qui suit, nous détaillons chaque phase à travers des exemples.

# 3. Processus de traduction

La phase de traduction consiste à traduire les EN déjà reconnues. Cette phase nécessite le passage par trois étapes : la traduction mot à mot, la réorganisation et l'accord et enfin le réajustement. Dans les sous sections suivantes, nous détaillons ces étapes.

## 3.1. Traduction mot à mot

Le principe de la traduction mot à mot est de traduire chaque mot composant l'EN c'est-à-dire effectuer une correspondance entre chaque mot de l'EN en langue source et son équivalent en langue cible à l'exception des mots qui n'existent pas dans les dictionnaires conçus ou qui ne peuvent pas être traduits (nombre, caractère spécial, etc.). Ceci est effectué en utilisant le transducteur de la **Figure 34**.

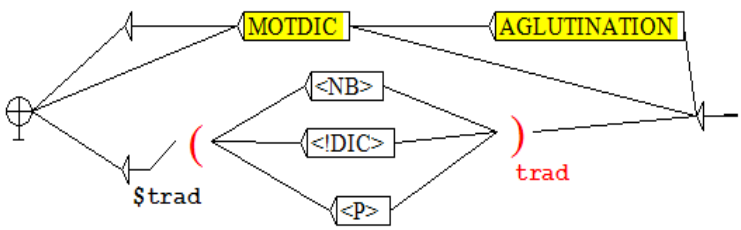


**Figure 34.** *Transducteur principal de la traduction mot à mot*

Le transducteur de la **Figure 34** prend comme entrée les EN déjà reconnues. Comme nous l'avons indiqué auparavant, ce transducteur tient compte, au sein d'une EN, des mots qui gardent les mêmes valeurs dans la langue cible. Ces mots peuvent être sous forme d'un nombre <NB> ou d'un caractère spécial <P> ou d'un mot qui n'existe pas dans les dictionnaires conçus <!DIC>.

Dans ce qui suit, nous présentons les différentes sous étapes formant la traduction mot à mot.

### 3.1.1. Traduction proprement dite

La traduction proprement dite consiste à traduire chaque mot de l'EN tout en précisant les différentes informations relatives aux noms qui sont étendus par des adjectifs. Ces informations sont données en langue source. La traduction proprement dite est décrite par le sous-graphe *MOTDIC*.

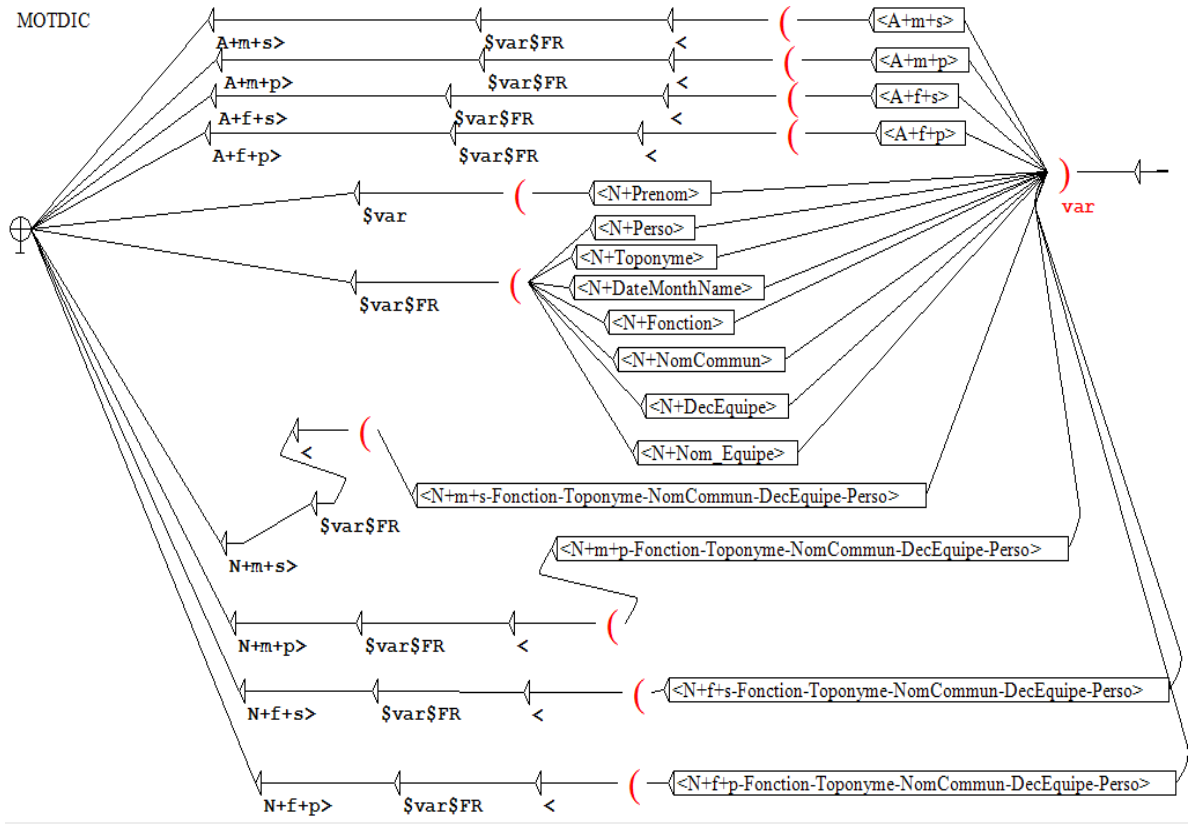


**Figure 35**. *Sous graphe "MOTDIC"*

Le transducteur MOTDIC donne comme résultat la traduction de chaque mot avec son annotation dans la langue source. Ces annotations sont attribuées à certaines catégories grammaticales (nom, adjectif) qui doivent être prises en compte afin d'effectuer l'accord et la réorganisation en langue cible dans une étape ultérieure. De plus, ces annotations portent des

informations de la langue source pour savoir à quel nom l'adjectif s'accorde. En effet, inversement à la langue française, dans la langue arabe l'adjectif n'accompagne pas obligatoirement le nom avec qui il s'accorde. Par exemple, si l'entrée à ce transducteur est l'EN مسبح مدينة الباسل الرياضية *masbaH madynat albaasil elriyaaDiyyah*, alors la sortie sera <piscine N+m+s> <cité N+f+s> <ville N+f+s> الباسل <sportif A+f+s>. Cela indique que le mot مسبح *masbah* est un nom (N), masculin (m) singulier (s) et sa traduction est piscine, le mot مدينة *madinat* est un nom féminin (f) singulier et peut avoir deux traductions cité et ville, le mot الباسل *el bacel* est un prénom (N+Prenom) donc il va rester dans la langue source et il va être translitteré ultérieurement et le mot الرياضية *elriyaaDiyyah* est un adjectif (A) féminin singulier et sa traduction est sportif. Le mot à traduire est sauvegardé dans la variable var et sa traduction est donnée par le mot technique $FR.

### 3.1.2. Élimination des traductions multiples

Les problèmes propres à la traduction mot à mot sont les traductions multiples qui peuvent être attribuées à un même mot. Pour résoudre ce problème, nous avons construit un transducteur qui permet d'éliminer les traductions multiples en fonction du contexte auquel est apparue l'EN. Le transducteur de la **Figure 36** permet par exemple d'éliminer l'ambigüité relative au nom مدينة *madinat* et au toponyme تونس *tounis*.

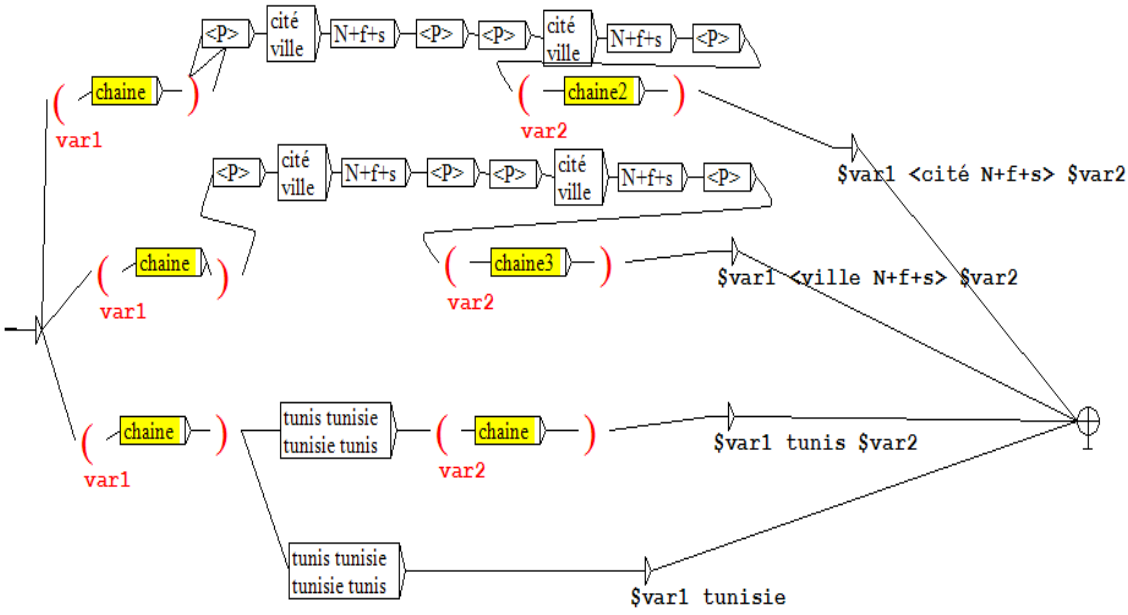


**Figure 36.** *Transducteur d'élimination des traductions multiples pour les mots* مدينة madinat et تونس tounis

Comme c'est déjà indiqué, le transducteur de la **Figure 36** traite le cas du nom مدينة *madynat* et du toponyme تونس *tuwnis*. Nous avons pris ces deux exemples car le premier admet des annotations (N+f+s puisque c'est un nom) données par l'étape précédente alors que le deuxième n'est pas annoté (puisque c'est un toponyme). Si nous reprenons la même EN que précédemment, nous remarquons que le mot مدينة *madynat* peut être traduit en cité et en ville. Dans ce cas, certains adjectifs décrits dans le sous graphe « chaine 2 » peuvent éliminer l'ambigüité. Par exemple, l'adjectif « sportif » peut nous indiquer que le mot «مـدينـة» doit être traduit en cité et non pas en ville. En effet, on dit cité sportive et non pas ville sportive. Donc, comme résultat de cette étape pour l'entrée <piscine N+m+s> <cité N+f+s> <ville N+f+s> الباسل <sportif A+f+s>, nous obtenons <piscine N+m+s> <cité N+f+s> الباسل <sportif A+f+s>.

Notons bien que l'élimination des traductions multiples pour les toponymes est plus délicate. Ce problème est trop fréquent pour la langue arabe. Dans ce cas, nous pouvons dans la plupart des cas éliminer l'ambiguïté suivant l'emplacement du toponyme.

## 3.2. Réorganisation et accord

Plusieurs règles de réorganisation et d'accord peuvent être déduites pour améliorer l'étape de traduction mot-à-mot. Ces règles ont essentiellement une relation avec l'ordre des mots qui constituent une EN. En effet, si une EN, dans la langue source, contient un adjectif alors il est nécessaire de savoir s'il appartient au mot déclencheur ou au nom qui vient juste avant lui. Par exemple, dans l'EN <piscine N+m+s> <cité N+f+s> الباسل <sportif A+f+s>, résultat de l'étape précédente, l'adjectif « sportif » est singulier et féminin (<sportif A+f+s>), le nom « cité » est aussi féminin singulier (<cité N+f+s>), mais le nom « piscine » est singulier masculin (<piscine N+m+s>). Nous déduisons alors que l'adjectif "sportif" appartient au nom « cité » et non au nom « piscine ».

Pour modéliser les règles de réorganisation et accord, nous avons construit le transducteur de la **Figure 37**. Ce transducteur s'applique autant de fois sur une même EN pour avoir, comme résultat, une EN traduite en langue cible réorganisée et accordée mais aussi sans aucune annotation.

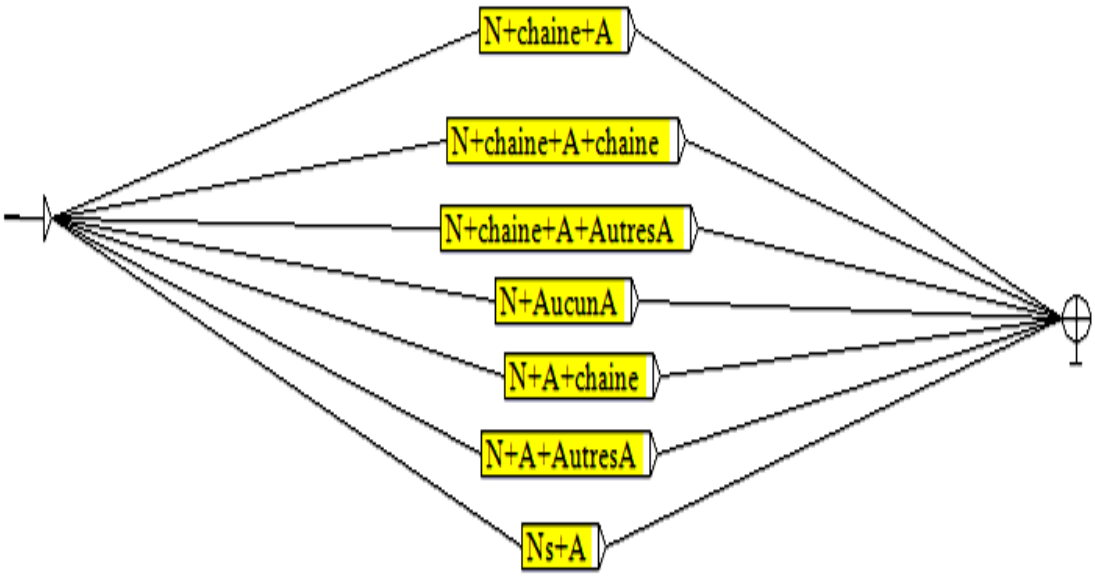


**Figure 37.** *Transducteur principal de réorganisation et accord*

Chaque sous-graphe du transducteur de la **Figure 37** traite un cas particulier pour la composition d'une EN traduite mot à mot. Cette EN peut avoir plusieurs formes :

- un nom annoté suivi d'une chaine de caractères suivi d'un seul adjectif annoté (N+chaine+A)
- un nom annoté suivi d'une chaine de caractères suivi d'un adjectif annoté suivi d'une autre chaine de caractères (N+chaine+A+chaine)
- un nom annoté suivi d'une chaine de caractères suivi d'un adjectif annoté suivi d'autres adjectifs annotés qui peuvent exister dans une chaine de caractères (N+chaine+A+AutresA)
- un seul nom annoté suivi d'un seul adjectif annoté suivi d'une chaine de caractères (N+A+chaine)
- un seul nom annoté suivi d'un adjectif annoté suivi par d'autres adjectifs (N+A+AutresA)
- plusieurs noms annotés suivis par un seul adjectif annoté (Ns+A)
- un nom annoté suivi d'un seul adjectif annoté (N+A).

L'application itérative du transducteur de la **Figure 37** à une même EN, traduite mot à mot, provoque le passage d'un sous graphe vers un autre jusqu'à n'avoir aucune annotation.

Dans l'exemple <piscine N+m+s> <cité N+f+s> الباسل <sportif A+f+s>, le transducteur de la **Figure 37** est appliqué deux fois. Dans la première itération, l'EN contient deux noms

annotées suivies par un seul adjectif annoté. Dans ce cas, le sous-graphe “Ns+A” est choisi comme chemin d’application. Ce sous-graphe est décrit dans la **Figure 38**.

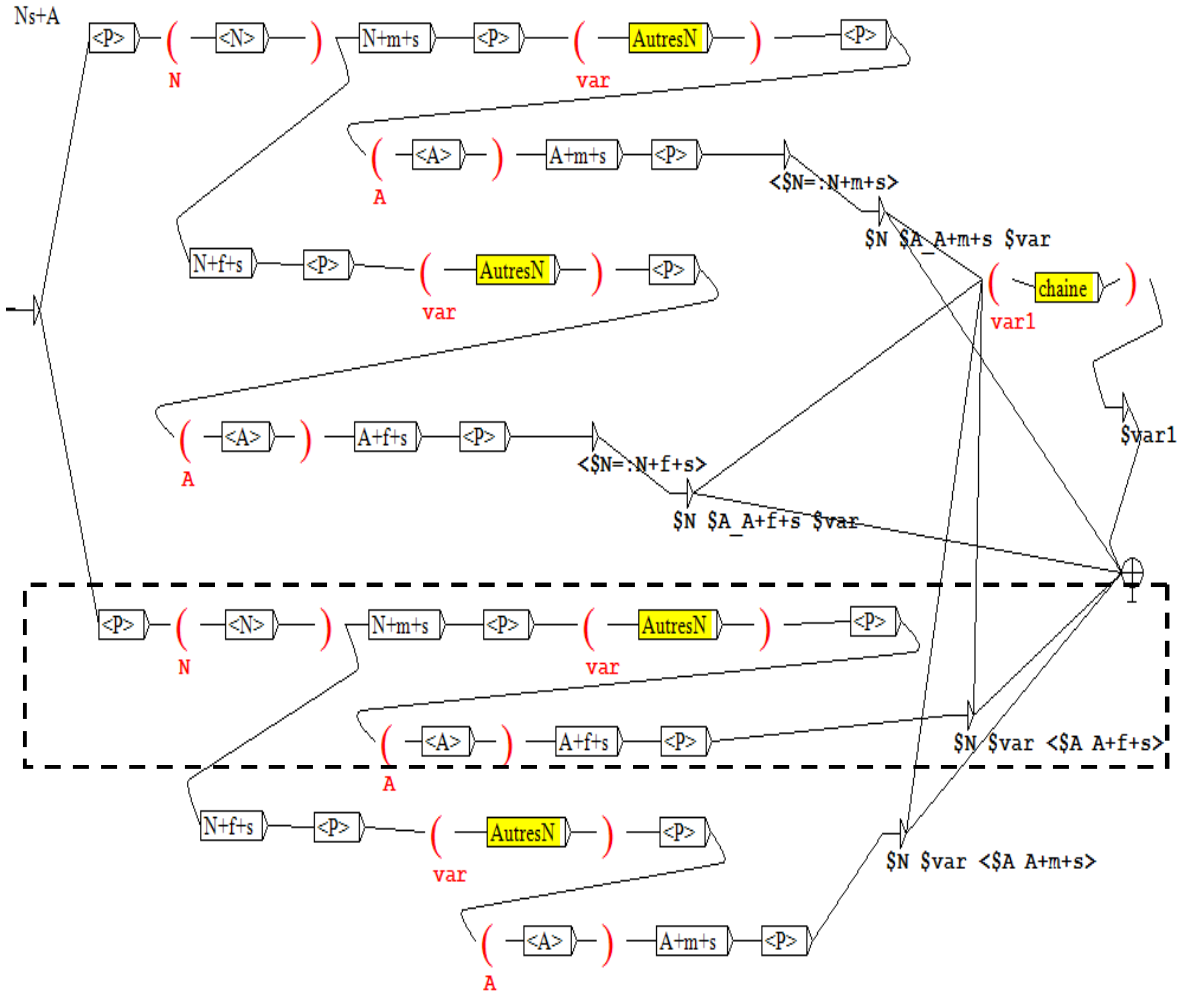


**Figure 38.** *Sous-graphe “Ns+A”*

Comme le montre la **Figure 38**, c’est le chemin encadré qui va être suivi. En effet, le premier nom annoté qui est piscine est masculin singulier en langue source (N+m+s) alors que l’adjectif est féminin singulier (A+f+s). Dans ce cas, le résultat obtenu est : piscine <cité N+f+s> الباسل <sportif A+f+s>. L’EN obtenue contient uniquement un seul nom annoté et un seul adjectif annoté. Le nom et l’adjectif sont séparés par d’autres mots. Par conséquence, dans la seconde itération, c’est le sous-graphe “N+chaine+A” qui va être suivi.

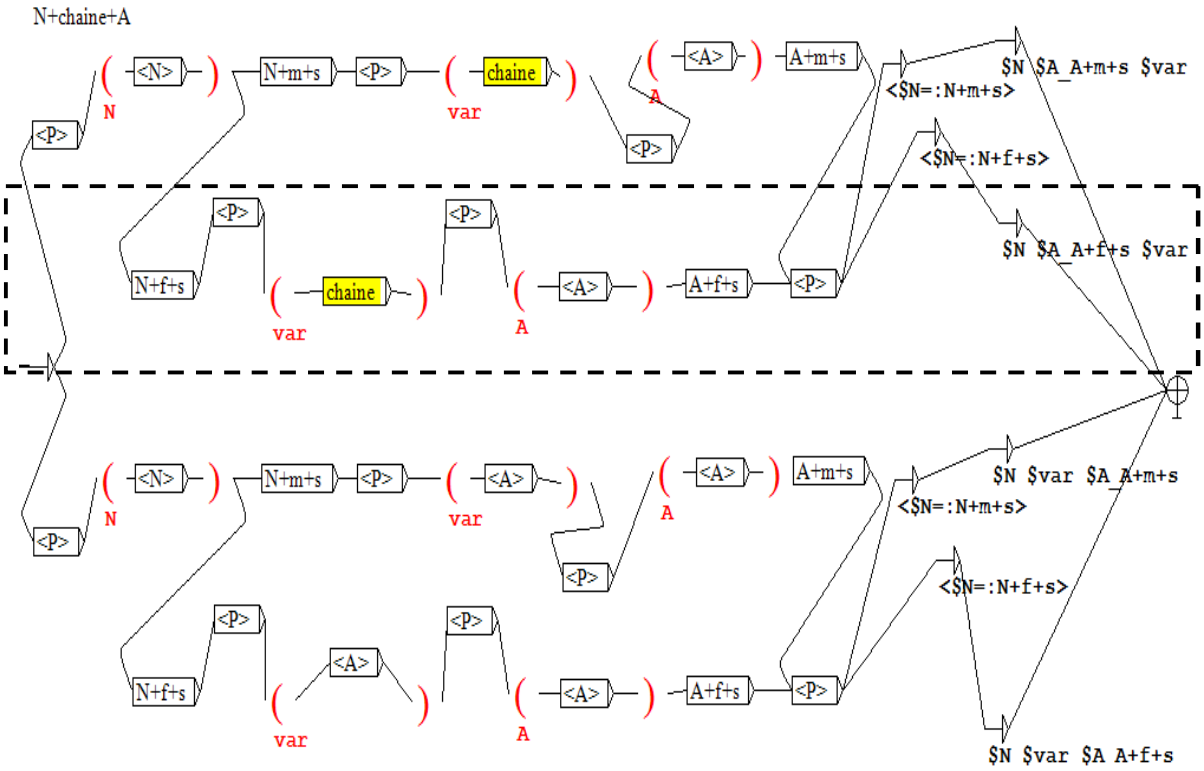


**Figure 39.** *Sous-graphe "N+chaine+A"*

Le chemin, encadré dans la **Figure 39**, est suivi dans le sous-graphe "N+chaine+A". Le passage par ce chemin donne comme résultat la traduction suivante : "piscine cité sportive الباسل". L'accord de l'adjectif avec le nom se fait tout d'abord en testant le genre et le nombre du nom. Dans notre cas, le test satisfait cette condition : <$N= : N+f+s>. Alors, l'adjectif doit suivre les mêmes attributs du nom et cela est effectué à l'aide de : $A_A+f+s. De plus, l'adjectif doit être déplacé juste après le nom avec qui il va s'accorder et ceci est effectué en utilisant : $N $A_A+f+s $var avec $N est le nom, $A_A+f+s est l'adjectif accordé et $var la chaine de caractères placée entre le nom et l'adjectif. Le résultat obtenu ne contient aucune annotation donc le traitement de l'EN concernée s'arrête pour cette étape.

## 3.3. Réajustement

L'étape de réajustement consiste à ajouter les prépositions dans les positions correspondantes de l'EN obtenue de l'étape de réorganisation et accord. Cet ajout est effectué si et seulement si l'EN contient des noms consécutifs ou bien un adjectif suivi par un nom. Par exemple, dans l'EN "piscine cité sportive الباسل", résultat de l'étape de réorganisation et accord, le nom «piscine» est suivi par un autre nom qui est «cité». Dans ce cas, il faut ajouter la préposition

«de la» entre « piscine » et « cité » car cité est un nom féminin singulier et commence par une consonne.

Les règles de réajustement sont modélisées avec le transducteur de la **Figure 40**.

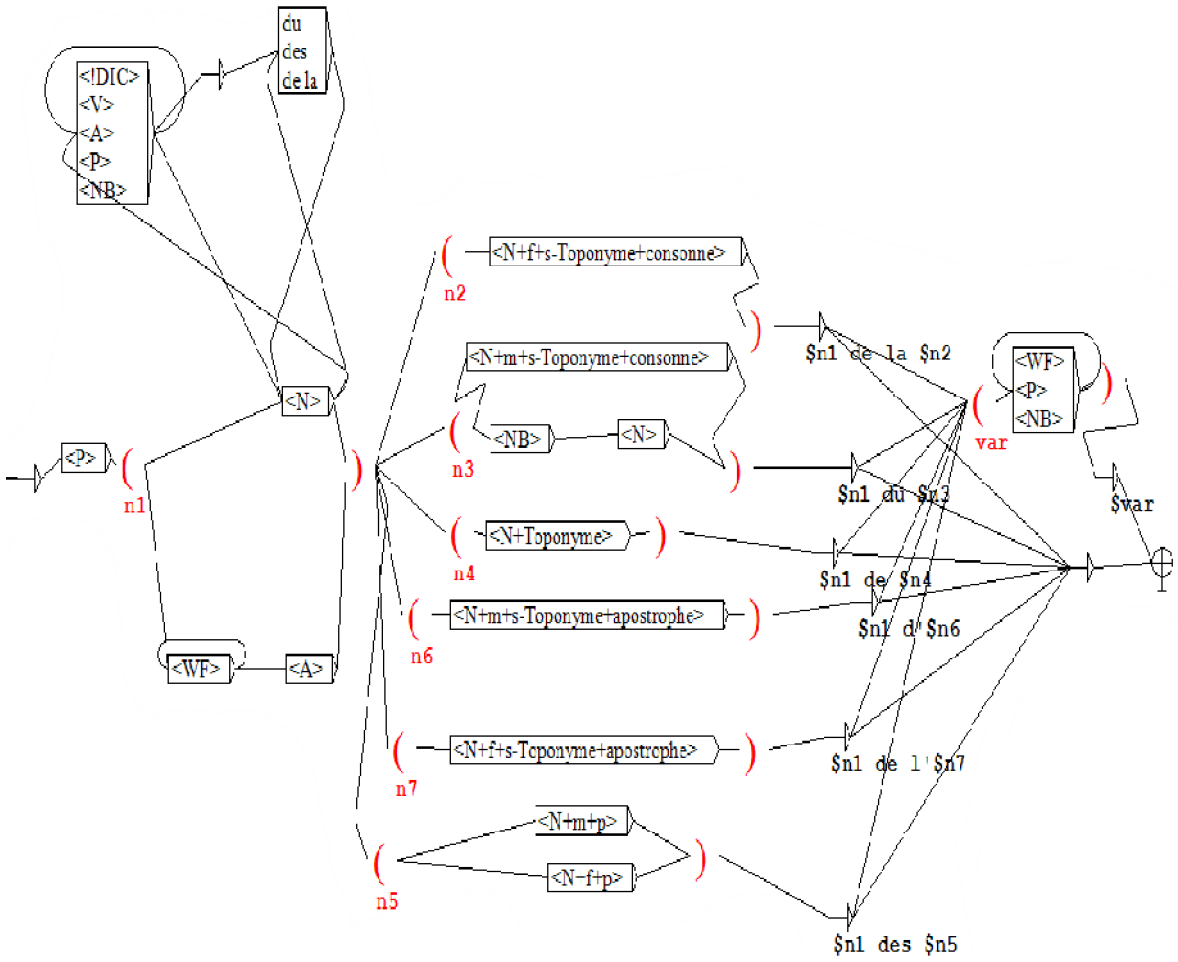


**Figure 40.** *Transducteur de réajustement*

Le transducteur de la **Figure 40** s'applique autant de fois sur une même EN jusqu'à traiter tous les noms consécutifs ou bien les adjectifs suivis par un nom. Le trait « apostrophe » dans ce transducteur indique que le nom commence par une voyelle et le trait « consonne » indique que le nom commence par une consonne. Cette condition est traitée par le transducteur de la **Figure 41**.

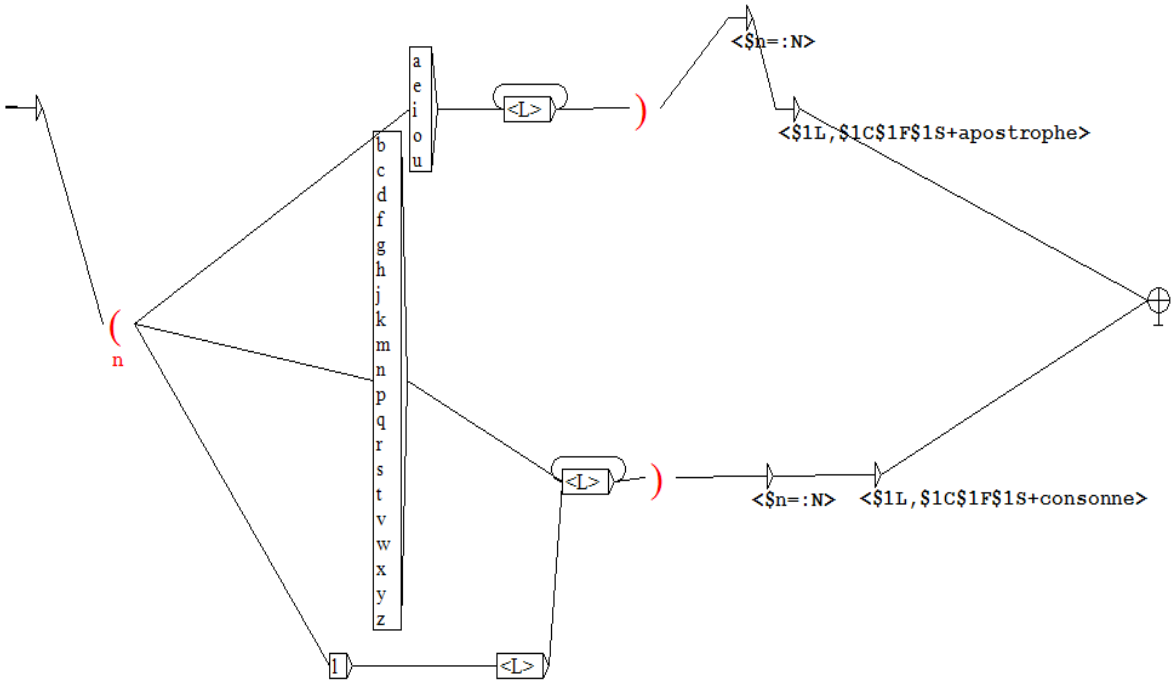


**Figure 41.** *Transducteur pour déterminer si un nom commence par une voyelle ou une consonne*

Dans le transducteur de la **Figure 41**, la lettre « l » est présentée dans un nœud à part (non avec les lettres consonnes) pour ne pas considérer le nom « la » (note de musique) qui peut générer dans l'étape de réajustement des conflits avec l'article défini « la ». Autrement dit, pour que la préposition « de la » ne sera pas ajoutée plus qu'une fois.

Le résultat final du processus de traduction pour l'exemple pris dans toutes les étapes est « piscine de la cité sportive الباسل ». Remarquons que dans cet exemple le mot « الباسل » reste toujours dans la langue source. En effet, il s'agit d'un prénom qui va être translitteré ultérieurement.

# 4. Processus de translittération

La translittération peut être vue comme la projection d'un mot d'une langue source vers une langue cible. Cette projection est fréquente dans toutes les langues vivantes. Elle permet une évolution rapide du vocabulaire de manière à s'adapter aux besoins des locuteurs. Une translittération peut reposer, d'une part, sur la correspondance phonologique ou morphologique et d'une autre part, sur la correspondance orthographique.

La translittération proprement dite est effectuée après avoir exécuté tous les transducteurs permettant la reconnaissance et la traduction des EN. Elle consiste à translittérer tous les mots non traduits qui sont écrits dans la langue source en utilisant les ressources convenables.

Dans le processus de translittération, nous tenons compte des règles de translittération qui respectent le système choisi Al-QALAM et des règles de transformation. Toutes ces règles sont modélisées avec des transducteurs morphologiques. La **Figure 42** représente un exemple de règle de translittération respectant le système Al-QALAM.

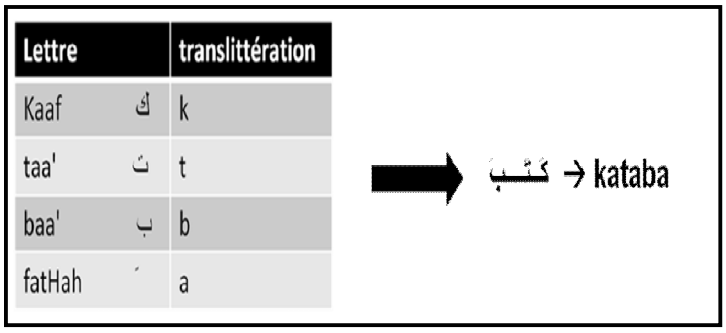


**Figure 42.** *Exemple de translittération du mot "كتب"*

Dans la **Figure 42**, la translittération du mot "كتب" consiste à donner l'équivalent de chaque lettre composant ce mot tout en respectant le système de translittération Al-QALAM. En appliquant cette correspondance entre lettres arabes et lettres françaises, nous devons prendre en compte les règles de transformation. Les figures 43 et 44 sont des exemples de ces règles.

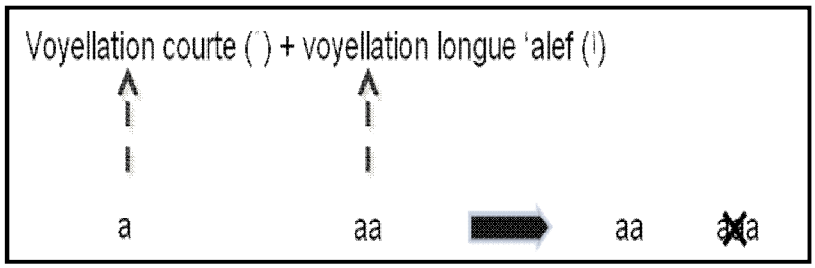


**Figure 43.** *Une règle de transformation concernant la voyellation longue*

La règle de la **Figure 43** indique s'il y a une voyelle courte comme "fatHah a" suivie par une voyelle longue comme "'alif aa", nous devons supprimer un "a". D'où au lieu d'obtenir "aaa", nous obtenons uniquement "aa".

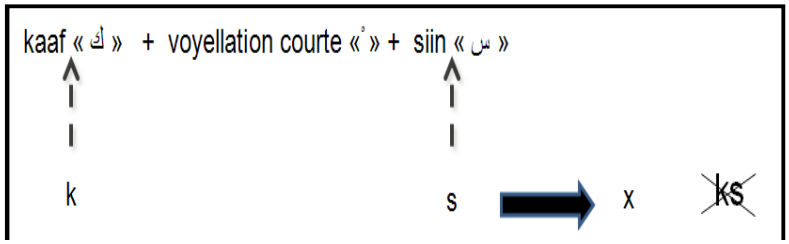


**Figure 44.** *Une règle de transformation concernant la voyellation*

La **Figure 44** illustre un autre exemple d'une règle de transformation. Cette règle indique s'il y a une lettre "kaaf k" voyellée par "sukuwn ْ" suivie par la lettre "syn s", alors nous devons transformer le résultat obtenu en "x". Par conséquent, au lieu d'obtenir "ks", nous obtenons "x". Par exemple, le nom propre "ماكس" doit être translitteré en "maax" et non en "maaks".

Toutes les règles de transformation et de translittération sont modélisées par le transducteur de la **Figure 45**.

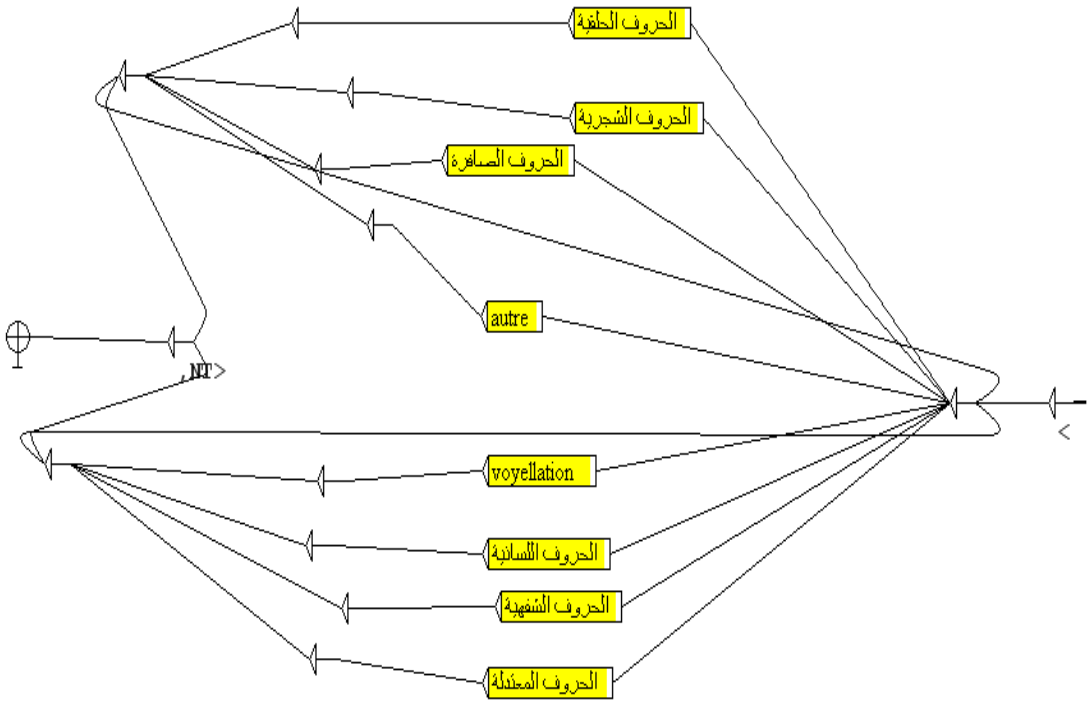


**Figure 45.** *Transducteur principal de translittération*

Le transducteur de la **Figure 45** représente toutes les classes des lettres arabes. Nous avons choisi de classifier les lettres arabes afin d'obtenir un transducteur plus lisible. Le transducteur représenté dans la **Figure 46** est un exemple d'un sous graphe du transducteur principal de translittération.

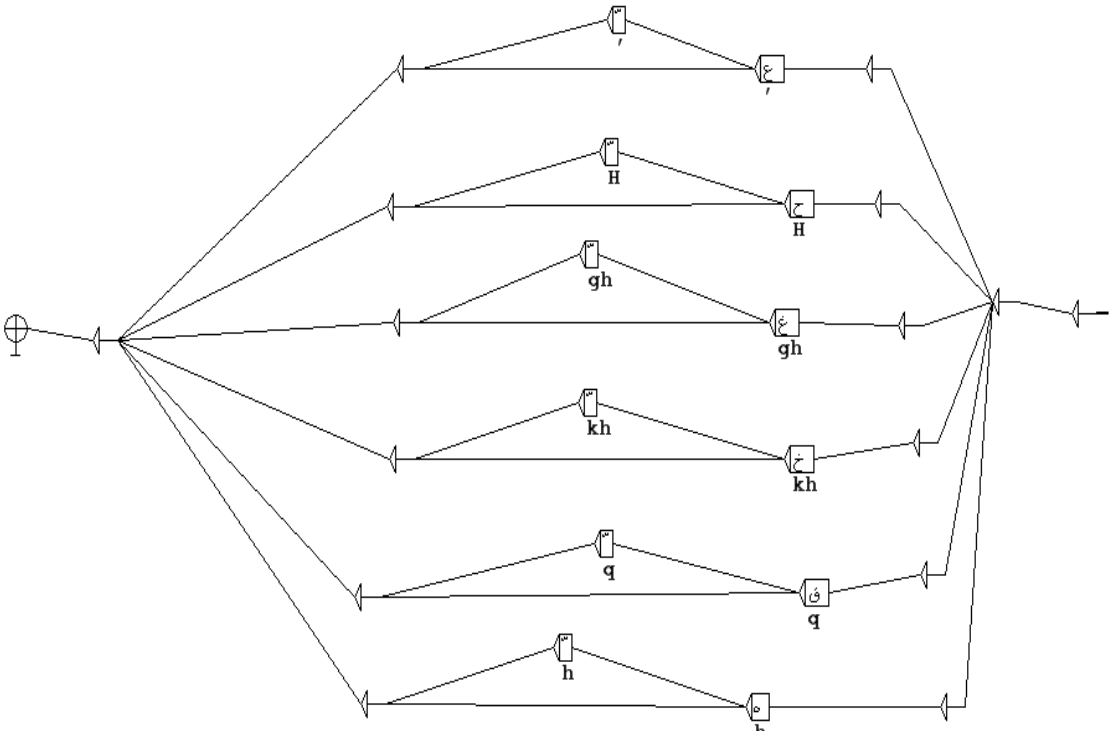


**Figure 46.** *Transducteur utilisant les règles de translittération et de transformation*

Le transducteur de la **Figure 46** décrit la translittération des lettres arabes classifiées : lettres de la gorge en leur équivalent en français. Rappelons qu'il faut voyeller le mot avant de le translittérer. Le transducteur permettant la voyellation des mots est représenté dans la **Figure 47**.

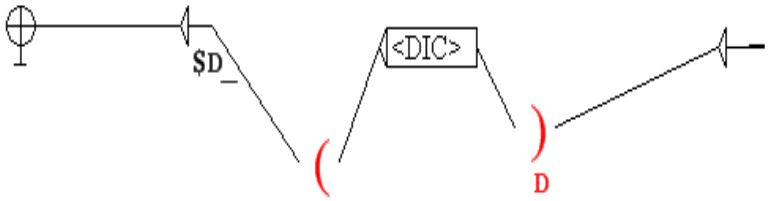


**Figure 47.** *Transducteur pour la voyellation*

Les outputs du transducteur de la **Figure 47** doivent être des inputs du transducteur représenté dans la **Figure 45**. En effet, le mot à translittérer passe initialement par la grammaire de voyellation pour être voyellé. Par la suite, le mot voyellé passe par le transducteur élaboré dans la **Figure 45** pour générer un mot bien translittéré.

# Conclusion

Dans ce chapitre, nous avons commencé par présenter certains problèmes qui doivent être pris en compte lors du processus de traduction. Ces problèmes nous ont permis de savoir qu'une traduction mot à mot n'est pas suffisante pour aboutir à développer un système de traduction robuste et cohérent. Ensuite, nous avons décrit les différentes étapes permettant une traduction qui respecte les spécificités de la langue française. Cette méthode de traduction va nous permettre de développer un système de traduction indépendant du module de la reconnaissance des EN mais aussi qui peut être réutilisable indépendamment du domaine. Nous avons encore décrit les différentes étapes pour l'intégration d'un module de translittération dans notre système de traduction. Tous les modules proposés soit pour la reconnaissance soit pour la traduction et y compris la translittération sont basés sur l'élaboration des grammaires morphologiques et syntaxiques. Ainsi, le choix d'une plateforme linguistique telle que NooJ, qui nous permet de tester ces grammaires, se révèle important. Cette expérimentation fait l'objet du chapitre suivant.

# Chapitre 6 : Mise en œuvre informatique avec l'environnement NooJ et évaluation

Après avoir détaillé la démarche proposée pour la reconnaissance des EN arabes et leur traduction vers le français dans les deux chapitres précédents, nous présentons dans ce chapitre l'outil réalisé afin d'expérimenter et valider cette démarche. En effet, nous avons choisi l'environnement de développement linguistique NooJ ainsi que son éditeur de commande noojapply pour mettre en œuvre l'outil de reconnaissance et de traduction des EN. NooJ est utilisé pour la construction des ressources linguistiques et la génération automatique des analyseurs morphologiques et syntaxiques. Afin d'appeler les dictionnaires et les transducteurs nécessaires pour chaque étape du processus et de garantir le fonctionnement récursif et automatique de ces étapes, nous avons conçu des scripts noojapply en C#.

L'évaluation des ressources déjà conçues est effectuée, pour la phase de reconnaissance, par l'utilisation des métriques d'évaluation : F-mesure, Rappel et Précision, d'une part, et par leur application sur un autre domaine que le sport, d'une autre part. Quant à l'évaluation de la phase de traduction, elle est effectuée à l'aide d'une étude comparative avec d'autres traducteurs.

Dans ce chapitre, nous commençons par donner un aperçu sur NooJ et sur noojapply. Ensuite, nous passons à la conception et à la modélisation de notre système de reconnaissance et de traduction. Puis, nous décrivons le fonctionnement de ce dernier à travers la description de quelques algorithmes. Enfin, nous détaillons et évaluons les résultats obtenus par notre système.

## 1. Implémentation NooJ

L'implémentation NooJ concerne la construction des ressources nécessaires soit pour la reconnaissance soit pour la traduction. Dans ce qui suit, nous donnons un aperçu sur NooJ et sur les ressources implémentées dans cette plateforme.

## 1.1. Aperçu sur NooJ / noojapply

Vu la compatibilité ascendante des outils et des formalismes de NooJ, qui sont graduellement plus puissants au fur et à mesure qu'on monte dans la hiérarchie linguistique, nous optons pour l'utilisation de cette plateforme. En effet, NooJ est considérée comme une plateforme linguistique de développement, un système de recherche documentaire, un extracteur terminologique, aussi bien que pour enseigner la linguistique et l'informatique linguistique aux étudiants (Silberztein & Tutin, 2005). Comme INTEX, cette plateforme est utilisée comme un outil de formalisation des langues naturelles et de développement d'applications de traitement automatique des langues naturelles (TALN).

NooJ est développée sous une architecture orientée objet (Silbeztein, 2004). De ce fait, les composants du système intègrent les données ainsi que les routines nécessaires pour leurs traitements. La communication et la coordination entre les objets sont réalisées par un mécanisme interne. Cette architecture se base souvent sur les trois piliers : l'encapsulation, l'héritage et le polymorphisme. L'architecture globale de NooJ est basée sur un ensemble de modules linguistiques tels que les modules orthographiques, flexionnels, morphologiques, dérivationnels et syntactico-sémantiques.

Les fonctionnalités de NooJ sont adaptées tout d'abord, à un public de linguistes qui s'intéressent à la description de la morphologie et de la syntaxe des langues, aux documentalistes s'intéressant à l'analyse de corpus, et enfin aux informaticiens du TAL pour les applications d'extraction d'information.

NooJ est un environnement linguistique de développement qui fournit des outils pour construire, tester et maintenir des descriptions formalisées à large couverture des langues naturelles ainsi que de développer des applications du TAL. Grâce à cette plateforme, nous avons la possibilité de construire des dictionnaires ainsi que des grammaires qui sont applicables aux textes afin de localiser les modèles morphologiques, lexicologiques et syntaxiques, enlever des ambiguïtés et étiqueter des mots simples et composés.

Dans son édition standard, les fonctions NooJ sont disponibles par l'intermédiaire d'un programme de ligne de commande : noojapply. noojapply peut être appelée soit directement à partir d'un "SHELL" script, ou de programmes plus sophistiqués écrits en Perl, C + +, Java, etc.

noojapply permet aux utilisateurs d'appliquer aux textes et aux corpus des dictionnaires et des grammaires automatiquement. Il peut être utilisé aussi dans un environnement professionnel

tel que la construction d'un moteur de recherche linguistique. Ceci est possible via une bibliothèque .NET dynamique : noojengine.dll, constituée par un ensemble de classes et de méthodes orientées objet publiques. Ces classes et méthodes peuvent être utilisées par n'importe quelle application. NET et dans tout langage de programmation .NET.

## 1.2. Implémentation des ressources

Comme nous l'avons déjà signalé, l'implémentation des ressources que nous avons identifiées et modélisées dans les deux chapitres précédents est effectuée avec la plateforme linguistique NooJ.

Concernant les dictionnaires, leurs entrées sont ajoutées manuellement et sont collectées à travers les sites web spécialisés (ex., www.kooora.com, www.koora.com et www.ar.wikipedia.org). Ces entrées sont représentées sous forme de lemmes avec la traduction française correspondante donnée par le mot technique «FR» et un modèle de flexion donné par le mot technique «FLX» pour éviter de représenter toutes les formes dérivées avec leurs traductions (ex., ملعب, N+Dec+FLX=MODEL1+FR=stade). La **Figure 48** est un extrait du dictionnaire des noms d'équipes.

```
#######################Equipes de Azerbaïdjan
ادليبى,N+Nom_Equipe+Azerbaïdjan+Football+FR=Adliyye
دينامو باكيلى,N+Nom_Equipe+Azerbaïdjan+Football+FR=Dinamo-Bakili
أف كى باكيلى,N+Nom_Equipe+Azerbaïdjan+Football+FR=FK Bakili
إف كي باكوا,N+Nom_Equipe+Azerbaïdjan+Football+FR=FK Baku
باكو,N+Nom_Equipe+Azerbaïdjan+Football+FR=Gomrukcu Baku
كاباز,N+Nom_Equipe+Azerbaïdjan+Football+FR=Kapaz G.
كاراباكا ازورزن,N+Nom_Equipe+Azerbaïdjan+Football+FR=Karabakh-Azersun
كارفان ايفلاك,N+Nom_Equipe+Azerbaïdjan+Football+FR=Karvan Evlakh
خازار لينكوران,N+Nom_Equipe+Azerbaïdjan+Football+FR=Khazar Lenkoran
كازر سامجيت,N+Nom_Equipe+Azerbaïdjan+Football+FR=Khazar Sumgayit
كازر يونفرسيتى,N+Nom_Equipe+Azerbaïdjan+Football+FR=Khazar Universiteti
لوكوميتيف ليمشلى,N+Nom_Equipe+Azerbaïdjan+Football+FR=Lokomotiv Imishli
ام كي تي - اراز,N+Nom_Equipe+Azerbaïdjan+Football+FR=MKT Araz
موبك باكو,N+Nom_Equipe+Azerbaïdjan+Football+FR=MOIK Baku
نيفتشي,N+Nom_Equipe+Azerbaïdjan+Football+FR=Neftchi
اولبيك باكو,N+Nom_Equipe+Azerbaïdjan+Football+FR=Olimpik Baku
شافا,N+Nom_Equipe+Azerbaïdjan+Football+FR=Shafa
شادوج سامرجوسار,N+Nom_Equipe+Azerbaïdjan+Football+FR=Shahdag-Samur Gusar
شامكير,N+Nom_Equipe+Azerbaïdjan+Football+FR=Shamkir
```

**Figure 48.** *Extrait du dictionnaire des noms d'équipes*

Comme décrit dans la **Figure 48**, toute entrée lexicale dans NooJ est constituée d'un ensemble de données, nous citons :

- un lemme : considéré comme forme de base ;
- une étiquette : qui en indiquera la catégorie grammaticale d'appartenance ;
- une liste optionnelle d'informations syntactico-sémantiques ;
- une liste éventuelle de codes alphanumériques désignant les modèles flexionnels et dérivationnels à y appliquer ;
- la traduction dans la langue cible.

Les dictionnaires ainsi établis doivent être compilés dans la plateforme NooJ avant d'être utilisés. Leur compilation est effectuée aussi dans NooJ.

Concernant les transducteurs, traitant et produisant des formes syntaxiques des EN, des traductions, des annotations et des transformations, nous n'allons pas les présenter ici puisque nous les avons déjà cités et expliqués dans les deux chapitres précédents. En plus de ces transducteurs, nous avons ajouté celui qui permet de résoudre le problème d'agglutination des EN arabes. Le principe de ce transducteur est de tester si le mot commence par ب bi, ف fa, ك ka, ل li, و wa. Si c'est le cas, alors on parcourt le reste du mot. Si la catégorie grammaticale du reste du mot est un nom ou un adjectif alors la sortie de cette grammaire sera l'annotation du premier alphabet composant le mot concerné par sa catégorie grammaticale correspondante : CONJ pour conjonction, PREP pour préposition et ARTDEF pour l'outil de détermination.

Le transducteur qui permet de traiter le problème d'agglutination se présente dans la **Figure 49**. Dans ce transducteur, la variable «harf» représente les prépositions qui peuvent être attachées à un nom ou à un adjectif. Ces prépositions sont : ب *bi*, ف *fa*, ك *ka* et ل *li*. La variable « def » contient l'article défini «ال *al*». Les prépositions peuvent précéder l'article défini et la conjonction «و *wa*» peut précéder les deux. De ce fait, si une préposition (ou une conjonction ou un article défini) est attachée à un ensemble de lettre (<L>) et si ces lettres constituent un nom ou un adjectif (<$Nom :=N> ou <$Nom= :A>) alors la préposition sera annotée par «PREP» (l'article défini par «ARTDEF» et la conjonction par «CONJ») et cette annotation va être utilisée ultérieurement dans d'autres grammaires qui font appel à la grammaire de la **Figure 49**.

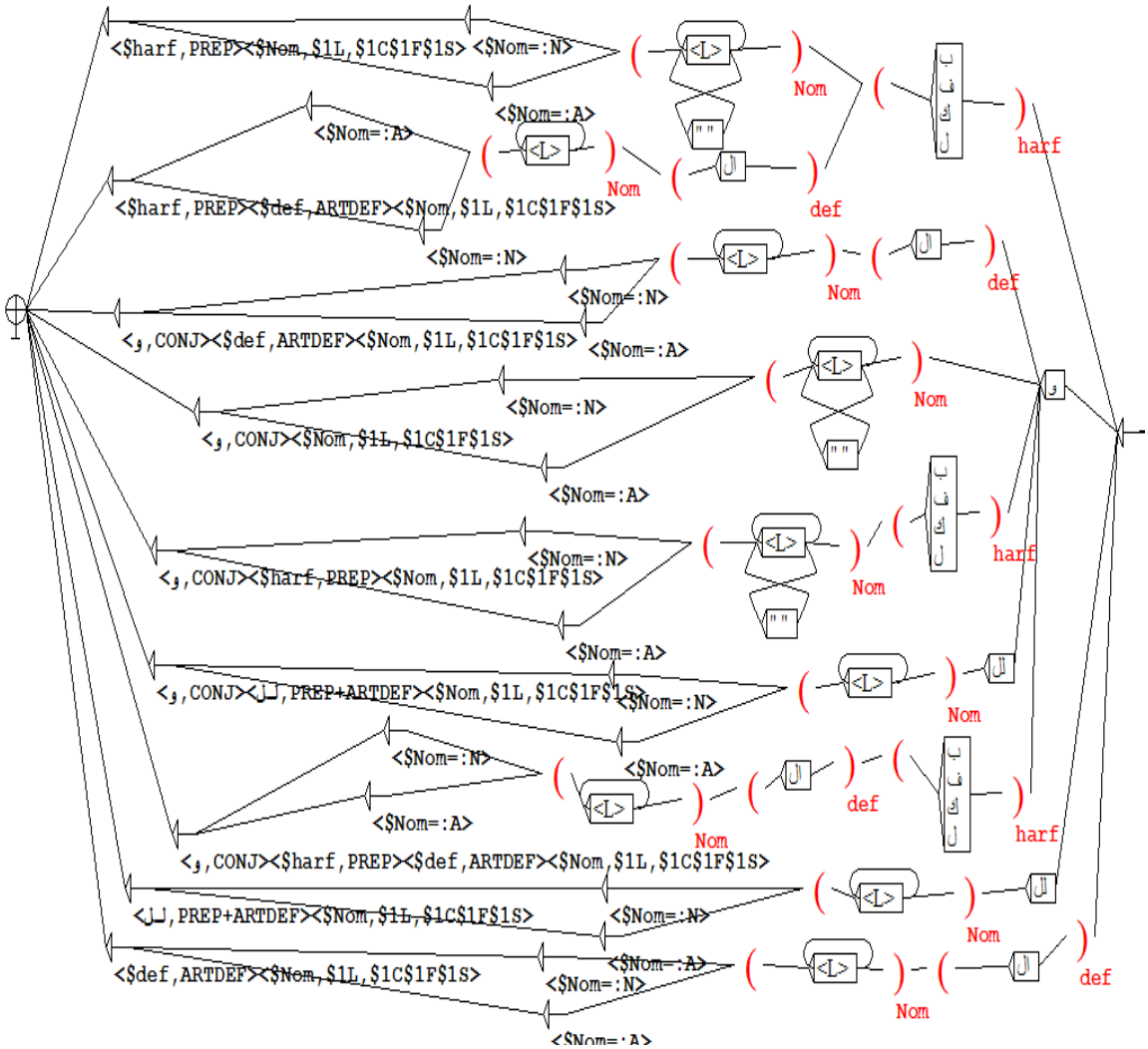


**Figure 49.** *Transducteur de résolution de l'agglutination*

Ce même transducteur prend en compte les noms composés tels que بحمّام = ب +حمّام الأنف الأنف (biHammam al'anf = bi + Hammam al'anf) et ceci à l'aide du nœud contenant l'espace (" "). Notons que les grammaires morphologiques utilisent les dictionnaires comme des entrées.

# 2. Implémentation C#

Comme nous l'avons déjà indiqué, les résultats de chaque étape du processus de reconnaissance et de traduction peuvent être utilisés comme des entrées pour l'étape qui suit. De plus, certaines étapes telles que la réorganisation et accord et le réajustement nécessitent une application itérative. Pour ces raisons, nous avons recours à l'utilisation des scripts noojapply implémentés en C# garantissant le bon fonctionnement de l'outil conçu.

Avant d'entamer l'implémentation de notre outil, nous décrivons dans la **Figure 50** les classes nécessaires pour la reconnaissance et la traduction des EN.

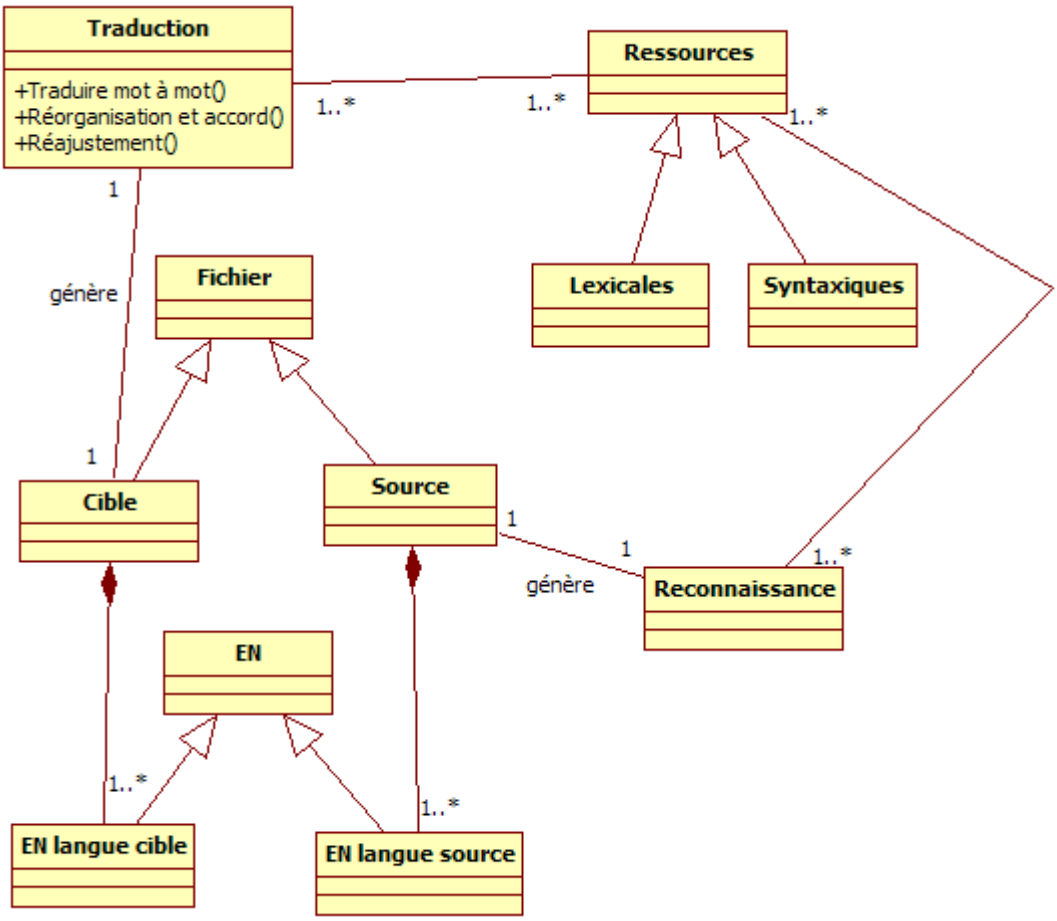


**Figure 50.** *Diagramme de classes de l'outil réalisé*

Dans la **Figure 50**, la classe «Traduction» constitue la classe la plus importante de notre prototype. En effet, cette classe contient toutes les méthodes permettant de réaliser la traduction des EN : *Traduire mot à mot()*, *Réorganisation et accord()* et *Réajustement()*. La classe «Traduction» se réfère à la classe «Ressources» qui contient toutes les ressources lexicales et syntaxiques identifiées dans la démarche proposée. Le fichier à traduire est composé d'une ou plusieurs EN dans la langue source. Quant au fichier résultat de la traduction, il est composé du même nombre d'EN mais dans la langue cible. Dans ce qui suit, nous présentons les algorithmes qui décrivent les différentes étapes de traduction : traduction mot à mot, accord et réorganisation, et réajustement.

## 2.1. Algorithme de traduction mot à mot

Rappelons que la fonction de traduction mot à mot consiste à traduire les constituants d'une EN sans tenir compte des spécificités de la langue cible. Les entrées de cette fonction sont : le fichier à traduire ($Fichier_{LS}$). La sortie de cette fonction est un fichier ($Fichier_{LC}$) contenant les

mêmes EN que dans $Fichier_{LS}$ mais traduites mot à mot. Le principe de fonctionnement de cette fonction est décrit par l'algorithme suivant :

---

**Entrée :** $Fichier_{LS}$

**Sortie :** $Fichier_{LC}$

**Algorithme** : $Traduction_{MàM}(Fichier_{LS,} Fichier_{LC})$

**Tant que** *(non (finfichier($Fichier_{LS}$))* **Faire**

chaine = ""

multiple =false

segmentation (ligne, L) /**/

Traduction $_{EN}$ (L, chaine, multiple)

**si** (multiple=true)

**alors**

CréerFichier(F, chaine)

EliminerTraductionMulitiples (Langue, F', D', $G_{syn}$, F)

CréerChaine(F',chaine)

**finsi**

CréerFichierFinal($Fichier_{LC}$, chaine)

Suivant(ligne)

**Fin Tant que**

---

Où :

- La fonction EliminerTraductionMultiples() permet d'éliminer les traductions multiples qui peuvent exister dans une EN. Ceci est effectué en utilisant le dictionnaire D' qui contient les mots d'une EN en français, la grammaire syntaxique décrite dans la **Figure 36** et le fichier F qui contient une EN traduite mot à mot.
- CréerChaine() permet de copier le contenu d'un fichier dans une variable de type chaine de caractères.
- CréerFichierFinal() permet d'ajouter une chaine dans un fichier.

La fonction de traduction mot à mot se sert de la fonction $Traduction_{EN}()$ qui permet de traduire les constituants d'une seule EN. Les entrées de $Traduction_{EN}()$ sont la liste des mots composant l'EN (L), la traduction de l'EN (chaine) et une variable indiquant si l'EN traduite contient des traductions multiples ou non. La fonction $Traduction_{EN}()$ est décrite par l'algorithme suivant :

---

**Entrée :** L

**Sortie :** chaine, multiple

**Algorithme** : Traduction $_{EN}$ (L, chaine, multiple)

**si** *(non (vide (L)))*

**alors**

CréerFichier (F, premier(L)) /*premier (L) retourne le premier mot de la liste L*/

CréerFichierTraductionNoojapply$_{Mot}$ (Langue, F', LD, $G_{Morph}$, $G_{syn}$, F)

ConstruireChaine(F', chaine, multiple)

Traduction $_{EN}$ (reste(L), chaine, multiple)

**finsi**

---

Où :

- La fonction CréerFichier() permet de copier le mot à traduire dans un fichier (F).
- La fonction CréerFichierTraductionNoojapply$_{Mot}$() qui permet de générer la traduction du mot concerné dans le fichier F'. Dans cette fonction, LD représente la liste de dictionnaires de la langue source tels que le dictionnaire des noms d'équipes, le dictionnaire de toponymes, $G_{morph}$ représente la grammaire d'agglutination décrite dans la **Figure 49** et $G_{syn}$ est la grammaire syntaxique représentée dans Figure **34**.
- La fonction ConstruireChaine() permet de construire la chaine (chaine) contenant la traduction de l'EN concernée à travers la concaténation de la traduction de chaque mot.

## 2.2. Algorithme de réorganisation et accord

La réorganisation et accord consiste à réorganiser les constituants des EN et assurer l'accord entre l'adjectif et le nom auquel il se rapporte. Cette fonction s'applique autant de fois sur les

différentes EN jusqu'à n'avoir aucune annotation. Les entrées pour cette fonction sont le fichier généré par la fonction de traduction mot à mot ($Fichier_{LC}$) et les ressources convenables. La sortie pour cette fonction est un fichier contenant des EN bien organisées et accordées ($Fichier_{Reorg}$). Le principe de fonctionnement de cette fonction est décrit par l'algorithme suivant :

---

**Entrée :** $Fichier_{LC}$

**Sortie :** $Fichier_{Reorg}$

**Algorithme** : ReorgEtAccord ($Fichier_{LC}$, $Fichier_{Reorg}$)

**Tant que** *(non (finfichier($Fichier_{LC}$))* **Faire**

chaine=ligne

**si** *(annotation(ligne))*

**alors**

ReorgEtAccord$_{EN}$ (chaine)

**finsi**

CréerFichierFinal($Fichier_{Reorg}$, chaine)

Suivant(ligne)

**Fin Tant que**

---

Où :

- La fonction annotation() indique si une EN contient des annotations ou non.
- La fonction ReorgEtAccord$_{EN}$() permet de réorganiser et accorder une seule EN. L'algorithme de cette fonction est le suivant :

---

**Entrée :** chaine /*EN à réorganiser et accorder*/

**Sortie :** chaine /*EN bien organisée et accordée*/

**Algorithme** : ReorgEtAccord$_{EN}$ (chaine)

CréerFichier(F, chaine)

CréerFichierNoojapplyReorg$_{EN}$(Langue, F', D', $G_{syn}$, F)

CréerChaine(F',chaine)

**si** *(annotation(chaine))*

**alors**

ReorgEtAccord$_{EN}$ (chaine)

**finsi**

---

Où :

- La fonction CréerFichierNoojapplyReorg$_{EN}$() permet de réorganiser et d'accorder pour une itération une EN. Ceci est effectué à l'aide de noojapply qui utilise les ressources D' (dictionnaire qui contient les composants d'une EN en français), $G_{syn}$ (grammaire de la **Figure 37**) et le fichier F.
- ConstruireFichierReorg(chaine, Fichier$_{Reorg}$) permet d'ajouter une chaine dans un fichier.

## 2.3. Algorithme de réajustement

Cette fonction permet l'ajout des prépositions convenables entre deux noms successifs ou après un adjectif suivi d'un nom. Les entrées de cette fonction sont le fichier contenant les EN réorganisées et accordées (Fichier$_{Reorg}$). La sortie pour cette fonction est un fichier contenant des EN qui respectent les spécificités de la langue cible (Fichier$_{Reaj}$). L'algorithme principal de cette fonction est le suivant :

---

**Entrée :** Fichier$_{Reorg}$

**Sortie :** Fichier$_{Reaj}$

**Algorithme** : Reajustement (Fichier$_{Reorg,}$ Fichier$_{Reaj}$)

**Tant que** *(non (finfichier(Fichier$_{Reorg}$)))* **Faire**

Reajustement$_{EN}$ (ligne)

CréerFichierFinal(Fichier$_{Reaj}$, ligne)

Suivant(ligne)

**Fin Tant que**

---

Où Reajustement$_{EN}$ () est une fonction qui permet de réajuster une seule EN (ligne). Cette fonction est décrite par l'algorithme suivant :

---

**Entrée :** chaine /*EN à réajuster*/

**Sortie :** chaine /*EN après réajustement*/

**Algorithme :** Reajustement$_{EN}$ (chaine)

CréerFichier(F, chaine)

CréerFichierNoojapplyReaj$_{EN}$(Langue, F', D', $G_{Morph}$, $G_{syn}$, F)

CréerChaine(F', res)

**si** *(non (vide(res)))*

**alors**

Reajustement$_{EN}$ (res)

**finsi**

---

Où CréerFichierNoojapplyReaj$_{EN}$() est une fonction qui permet de réajuster pour une itération une EN. Les entrées pour cette fonction sont le dictionnaire D' (même dictionnaire utilisé dans la fonction de réorganisation et accord), une grammaire morphologique $G_{Morph}$ (celle décrite dans la **Figure 41**), une grammaire syntaxique $G_{syn}$ (celle décrite dans la **Figure 40**) et un fichier F contenant une EN. Le résultat donné par cette fonction est sauvegardé dans F'.

Notons que la fonction CréerFichier(), qui permet la copie d'une EN dans un fichier, est obligatoire car noojapply nécessite que l'argument soit un fichier et non une chaine de caractères.

Après avoir donné une idée sur les algorithmes élaborés, nous présentons ci-dessous l'expérimentation et l'évaluation de l'outil.

# 3. Expérimentation et évaluation

L'expérimentation de l'outil de reconnaissance et de traduction des EN arabes établi est effectuée dans la plateforme linguistique NooJ. Comme nous l'avons déjà signalé, cette plateforme utilise les grammaires (syntaxiques et morphologiques) et les dictionnaires déjà conçus et édités dans NooJ. Le **Tableau 8** ci-dessous donne une idée sur les dictionnaires ajoutés aux ressources de NooJ.

| Dictionnaire | Nombre d'entrées | Annotation dans le dictionnaire |
|---|---|---|
| **Noms de joueurs** | 18000 | N+Joueur |
| **Noms d'équipes** | 5785 | N+Equipe |
| **Noms de sports** | 337 | N+Sport |
| **Noms de pays et capitales** | 610 | N+Toponyme |
| **Noms de personnalités** | 300 | N+Perso |
| **Mots déclencheurs** | 20 | N+Dec |

| | | |
|---|---|---|
| **Noms des fonctions** | 100 | N+Fonction |

**Tableau 8.** *Dictionnaires ajoutés aux ressources de la plateforme NooJ*

Aux dictionnaires mentionnés dans le **Tableau 8**, l'outil conçu utilise d'autres dictionnaires NooJ tels que les dictionnaires des adjectifs, des noms, des prénoms (Mesfar, 2008). A ces dictionnaires, nous avons ajouté quelques entrées relatives au domaine du sport et la traduction en français de ses entrées. Notons que le dictionnaire de prénom reste monolingue car ses entrées peuvent être translittérées.

L'expérimentation de notre outil est effectuée sur deux domaines : le domaine du sport et le domaine d'enseignement. Le premier est choisi car il est le sujet de notre corpus d'étude. Quant au deuxième domaine, il est choisi pour prouver que notre outil peut être appliqué sur d'autres domaines. Dans la sous section suivante, nous expérimentons et évaluons tout d'abord la phase de reconnaissance et ensuite la phase de traduction.

## 3.1. Expérimentation de la phase de reconnaissance

Pour évaluer la phase de reconnaissance, nous avons appliqué notre outil sur un corpus différent de celui utilisé lors de l'étude des EN. Ce corpus est formé de 4000 textes du domaine du sport environ 94,5Mo collectés des différents journaux quotidiens (ex., assabah, alanwar, el chourou9, al ahram) et de wikipédia. Le corpus contient 180000 EN appartenant aux différentes catégories du domaine du sport (ex., nom de joueur, nom de sport, terme sportif). De ces EN, il y a 40000 EN appartenant à la catégorie *Nom de lieu* (c.-à-d., stade, salle, cité, piscine et complexe). Ces EN sont recensées manuellement et à l'aide des requêtes NooJ. Nous estimons que cette taille de corpus est assez représentative du domaine eu égard à la diversité géographique des sources et la diversité des spécialités. Le corpus obtenu est ensuite nettoyé en éliminant en particulier les images qu'il peut contenir. Cela permet d'alléger notre corpus et rendre son chargement en mémoire plus rapide pour les différents tests.

Les résultats obtenus sont illustrés dans la table de concordances de la **Figure 51** générée par la plateforme NooJ.

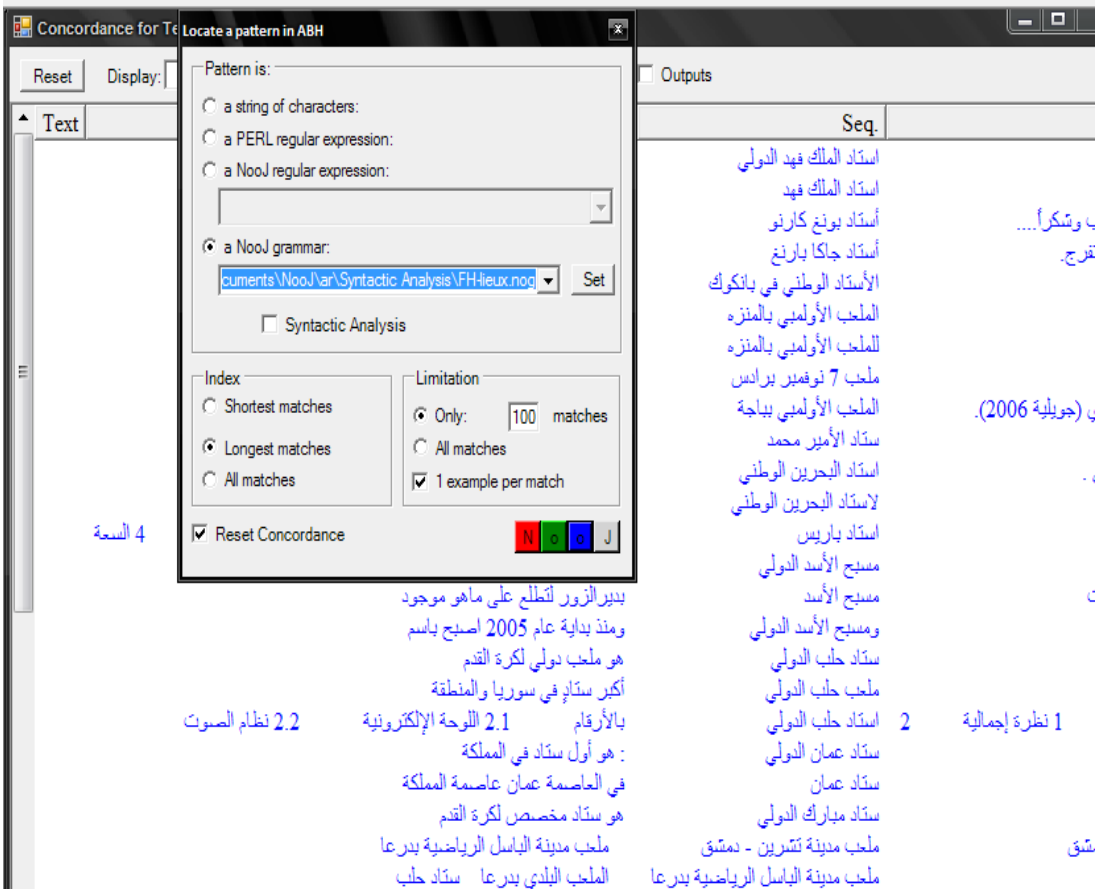


**Figure 51.** *Table de Concordances des EN reconnues dans le domaine du sport*

Comme il est indiqué dans la **Figure 51**, la table de concordances contient les différentes EN reconnues telles que الملعب الأولمبي بالمنزه *almal`ab al'uwlimby bilmanzah* (*Stade olympique el-Manzeh)*, ستاد حلب الدولي staade Halab elduwaly (*satde international de Alep)*. Notons qu'une EN est détectée si elle vérifie l'un des chemins décrit par le transducteur de la **Figure 28** du chapitre 4. En effet, un transducteur est caractérisé par un seul nœud de début et plusieurs nœuds de fin. Si plusieurs chemins sont vérifiés, le chemin le plus long est pris par défaut (il s'agit d'une option dans la plateforme NooJ).

Les résultats de la **Figure 51** sont interprétés par les calculs de métriques suivantes : la précision, le rappel et la F-mesure ($F_1$). Le **Tableau 9** illustre les calculs des métriques obtenus.

| | **Nbre d'occurrences (Lieux sportifs)** | **Précision** | **Rappel** | **F-mesure** |
|---|---|---|---|---|
| **Textes journalistiques (domaine de sport) 4000 textes (94,5 Mo)** | 40000 | 0,98 | 0,90 | 0,94 |

**Tableau 9.** *Résultats obtenus*

Les valeurs des mesures du **Tableau 9** montrent l'existence de quelques problèmes non résolus. Certains de ces problèmes sont liés à l'absence de normes pour l'écriture des noms propres (ex. el hamza), à l'absence de quelques mots dans les dictionnaires construits (ex., mots provenant du dialecte tunisien), à la présence des unités contenant des caractères de décoration «مـــلـــعـــب» et au manque du blanc entre deux termes tel que le mot composé «حمام الأنف». Ceci provoque un silence. D'autres problèmes sont liés à des concepts spécifiques à la langue arabe comme la métaphore. Par exemple, nous pouvons trouver dans un texte une EN composée d'un mot déclencheur appartenant au domaine du sport suivi d'un nom d'une personne célèbre suivi d'une ville tel que ملعب الأسد بسوريا *mal`ab al'asad bisuwriya* ***stade el Assad à Syrie***. Le contexte dans lequel est apparu cette EN n'a pas l'objectif de citer un nom de stade mais de montrer les compétences du roi el Assad dans le sujet cité. Ce genre de problèmes est rare mais cause du bruit.

Nous avons aussi appliqué l'outil sur un corpus de 300 textes (environ 14.5 Mo) contenant 3000 institutions universitaires et collectés des différents sites web et de wikipédia. Comme résultat, nous avons obtenu la table de concordances suivante.


**Figure 52.** *Table de concordances concernant les institutions universitaires*

Les résultats obtenus sont interprétés par les valeurs de métriques choisies : 98% de précision, 70% de rappel et 82% de F-mesure. Nous remarquons que le silence est augmenté. Cela est dû à l'incomplétude des dictionnaires spécifiques à ce domaine (ex., les noms de personnalités dans le domaine du sport ne sont pas les mêmes que dans le domaine de l'enseignement, les noms de villages dans le domaine d'enseignement sont nombreux par rapport au domaine du sport) vu que ces dictionnaires sont en cours de construction.

Aussi, l'outil peut être appliqué indépendamment du domaine à condition que nous utilisions les mêmes traits adoptés dans les dictionnaires que nous avons construits. Il est évident que pour des raisons liées au domaine choisi, nous devons parfois ajouter d'autres chemins et d'autres sous-graphes aux graphes déjà construits. Cependant, nous ne sommes pas menés à reconstruire toutes les grammaires dès le début.

Les EN arabes obtenues lors de la phase de reconnaissance vont êtres sauvegardées dans un fichier ayant l'extension .ind pour être exploitées dans la phase de traduction. Ce fichier est généré par l'éditeur de commande noojapply en utilisant l'instruction suivante : noojapply (la langue, le nom du fichier résultat (.ind), la liste des dictionnaires (.nod), la liste des grammaires morphologiques (.nom), la liste des grammaires syntaxiques (.nog), le corpus (.noc)). Il est à noter que l'application des différentes ressources au corpus que ce soit par noojapply ou dans NooJ permet l'annotation des différents textes composant le corpus.

## 3.2. Expérimentation de la phase de traduction

Le processus de traduction est appliqué sur les EN arabes obtenues lors du processus de reconnaissance. Notons que les résultats erronés sont hérités. C'est pourquoi des heuristiques de filtrage semblent nécessaires avant le passage à la traduction. Cette phase comme ci-indiqué dans le chapitre précédent passe par trois étapes. Les résultats obtenus dans l'étape de traduction mot à mot sont illustrés dans la **Figure 53**.

```
725,747,<stade N+m+s> roi فهد  <international A+m+s>
6929,6944,<stade N+m+s> roi فهد
7308,7324,<stade N+m+s> Bung Karno
7402,7418,<stade N+m+s> Jaka Baring
8503,8526,<stade N+m+s> <olympique A+m+s> El Menzah
8884,8903,<stade N+m+s> 7 novembre Radès
9118,9139,<stade N+m+s> <olympique A+m+s> Beja
12187,12202,<stade N+m+s> Alexandria
12743,12759,<stade N+m+s> Ismaily
13000,13016,<stade N+m+s> prince محمد
13202,13222,<stade N+m+s> Bahreïn <national A+m+s>
13807,13817,<stade N+m+s> As-Salt
15399,15414,<stade N+m+s> Borg El Arab
15617,15630,<stade N+m+s> El-Mansora
16496,16507,<stade N+m+s> Paris
18561,18575,<stade N+m+s> Borg El Arab
20095,20107,<stade N+m+s> Port Said
20353,20368,<stade N+m+s> <université N+f+s> Beijing
22585,22612,<piscine N+m+s> الأسد  <international A+m+s> Dayr az-Zur
23501,23518,<piscine N+m+s> الأسد  <international A+m+s>
28347,28362,<stade N+m+s> Haras Alhodod
28801,28816,<stade N+m+s> Alep <international A+m+s>
32846,32862,<stade N+m+s> Amman <international A+m+s>
33501,33518,<stade N+m+s> مبارك  <international A+m+s>
33676,33692,<stade N+m+s> Mohammed Al-Hamad
34847,34866,<stade N+m+s> Alfayhaa - Damas
34895,34913,<stade N+m+s> libération - Damas
34917,34940,<stade N+m+s> <ville N+f+s> Techrine - Damas
34944,34976,<stade N+m+s> <cité N+f+s> الباسل  <sportif A+f+s> Daraa
34980,34999,<stade N+m+s> <municipal A+m+s> Daraa
35090,35104,<stade N+m+s> Al-Hamdaniya
35114,35139,<stade N+m+s> Khaled ibn El Walid - homs
```

**Figure 53.** *Extrait des résultats de la traduction mot-à-mot*

Comme le montre la **Figure 53**, le fichier obtenu contient les même EN que le fichier des EN reconnues mais qui sont traduites avec annotation. Cette traduction ne tient pas compte de l'accord et des spécificités de la langue française. Ainsi, l'application des ressources qui permettent l'accord de l'adjectif avec le nom auquel il est associé et la réorganisation des constituants de chaque EN donne comme résultat le fichier représenté dans la **Figure 54**.

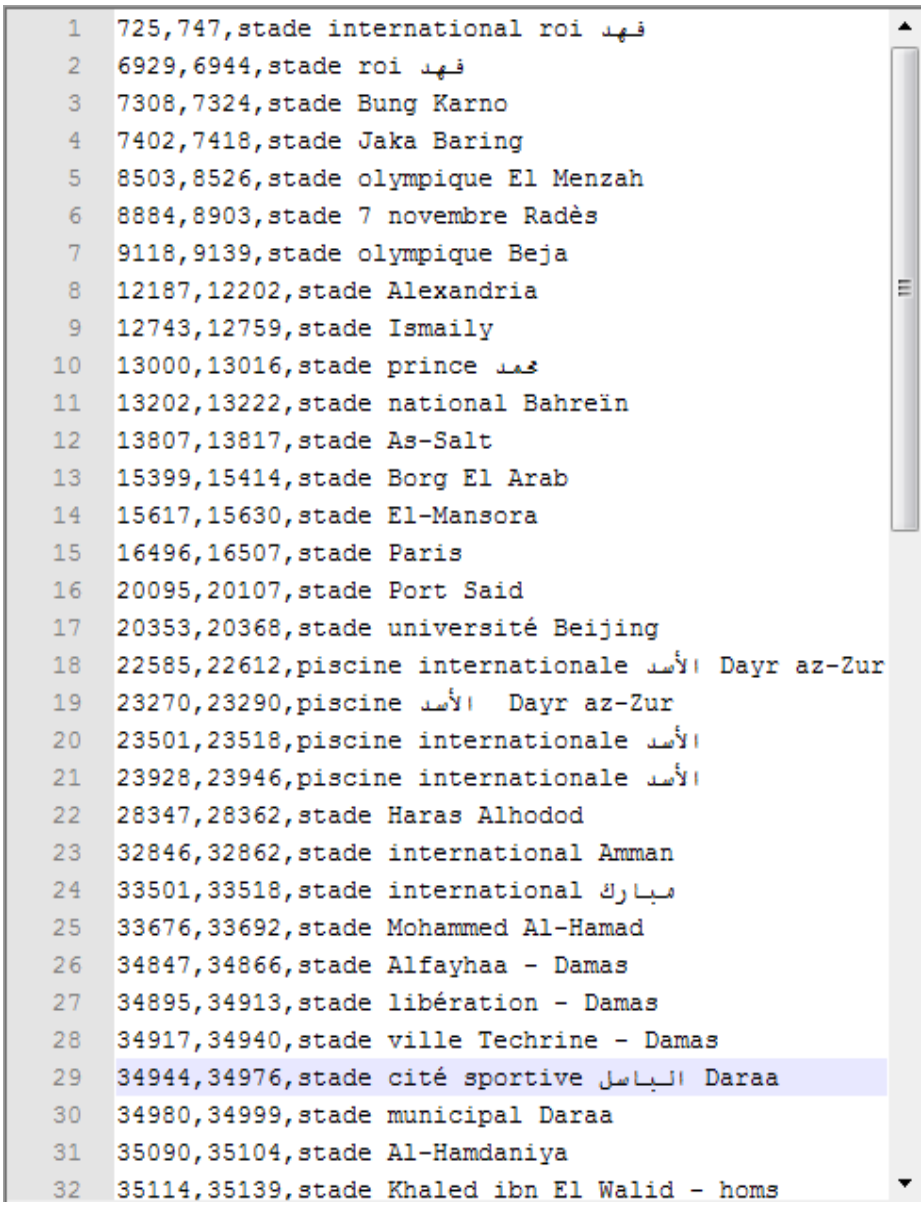


**Figure 54.** *Extrait des résultats de réorganisation et accord*

Les EN décrites dans le fichier de la **Figure 54** sont bien organisées et chaque adjectif s'accorde en genre et en nombre avec le nom auquel il se rapporte. Ceci grâce aux annotations générées par l'étape de traduction mot à mot comme ci-indiqué dans le fichier de la **Figure 53**.

Afin d'avoir une traduction plus fine, nous avons appliqué des règles de réajustement. L'application de ces règles donne comme résultat le fichier illustré dans la **Figure 55**.

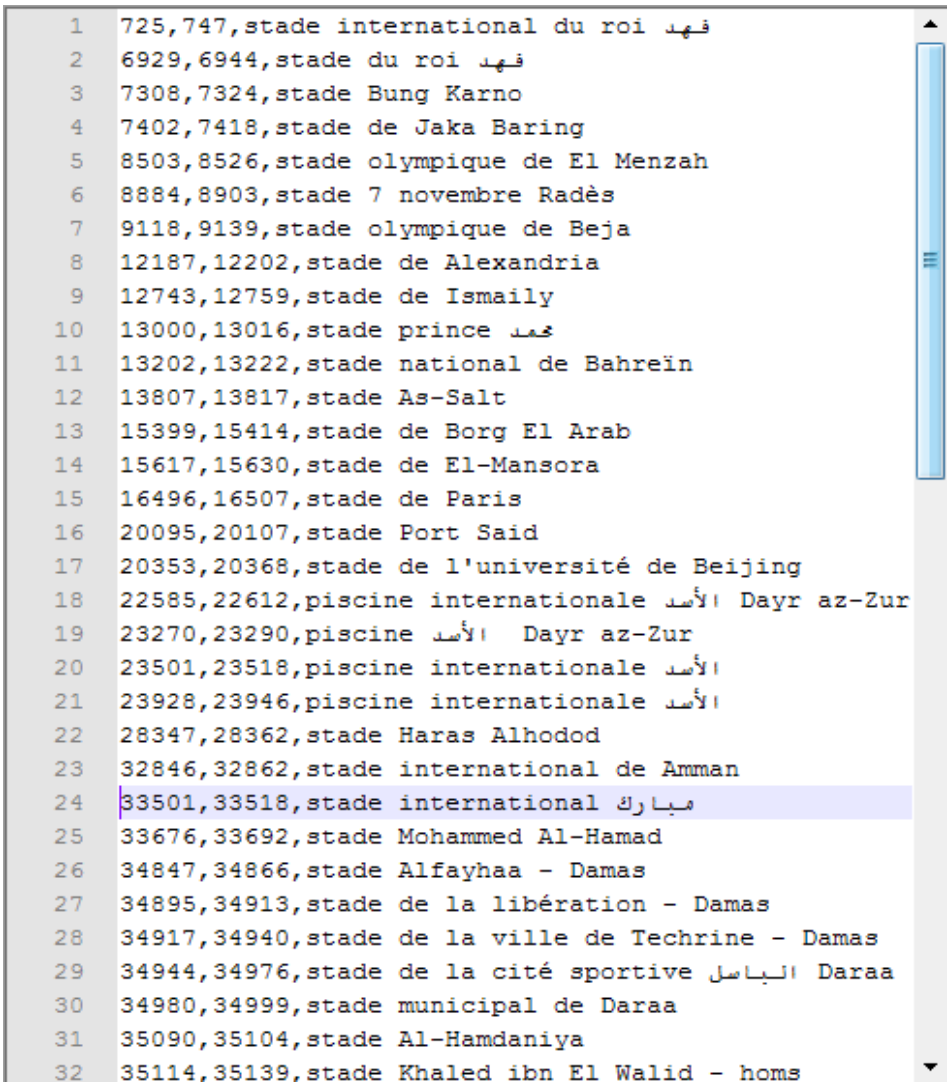


**Figure 55.** *Extrait des résultats de réajustement*

Comme le montre l'extrait du fichier de la **Figure 55**, le réajustement est effectué à travers l'ajout des prépositions tel que «du» avant le mot «roi» et «de l'» avant université. Remarquons que dans ce fichier, il y a des mots qui ont resté dans la langue source tels que فهد *fahd* et الأسد *al'asad*. Ces mots vont être translittérés en appliquant les ressources correspondantes et déjà décrites dans le chapitre précédent. Ceci génère le résultat final illustré dans l'extrait du fichier de la **Figure 56**. Remarquons que dans ce fichier, il n'existe aucun constituant de l'EN dans la langue source. Ce fichier représente les EN obtenues dans la phase finale. Ces EN sont bien traduites et translittérées selon le système al-Qalam.


**Figure 56.** *Extrait du fichier résultat après translittération*

Notre outil de traduction fournit 97% des EN bien traduits (c.-à-d., garantissant les spécificités de la langue cible). Le résultat obtenu est satisfaisant et démontre qu'il y a quelques problèmes non résolus. Ces problèmes sont liés aux traductions multiples affectées à un même toponyme (ex.,*tuwnis* تونس peut être traduit en tunisie ou tunis).

Nous avons comparé notre outil de traduction avec d'autres traducteurs connus tels que Google et Babylon qui supportent la traduction de l'Arabe vers le Français.  Les résultats obtenus sont illustrés dans le **Tableau 10**.

| | ENA | L'outil développé | Google | Babylon |
|---|---|---|---|---|
| a | مسبح مدينة الباسل الرياضية | stade de la cité sportive Albaasil | Bâle sport de la ville piscine | Une ville courageux sportives |
| b | الملعب الأولمبي بحمام الأنف | stade olympique de hammam lîf | Nez olympique de natation Stade | Solidarité susmentionnés stade olympique |
| c | استاد الملك فهد الدولي | stade international du roi Fahd | King Fahd International Stadium | stade le Roi Fahd internationale |
| d | ملعب مدينة تشرين - دمشق | stade de la ville de Techrine - Damas | City Stadium Octobre - Damas | aire de jeu ville Octobre-Damas |
| e | ملعب 7 نيسان - القامشلي | stade 7 nisan - Kamishly | Stade Avril 7 - Qamishli | 7 avril aire de jeu-القامشلى |
| f | ملعب خالد بن الوليد - حمص | stade Khaled ibn El Walid - homs | Khaled Bin Al Waleed Stadium - Homs | aire de jeu Khalid Bin Al-Walid-A |

**Tableau 10.** *Résultats expérimentaux*

Comme le montre le **Tableau 10**, les deux traducteurs Google et Babylon donnent des résultats incorrects dans tous les exemples cités à l'inverse de notre système. Par exemple, pour l'EN مسبح مدينة الباسل الرياضية (**Tableau 10**(a)) *masbah madinat al bacel al riadhiya*, notre outil donne comme résultat *piscine de la cité sportive Albaasil* qui est bien traduite et qui respecte les spécificités de la langue cible. Cependant, Google génère *Bâle sport de la ville piscine* qui est inconsistante et qui modifie le sens. De même Babylon donne *Une ville courageux sportives* qui est n'est pas correcte.

Notons aussi qu'il y a des EN qui sont traduites par Google et par Babylon vers l'anglais au lieu du français (**Tableau 10** (c et d)). En outre, les noms composés ne sont pas traités ni par Google ni par Babylon (**Tableau 10** (b)). Au contraire, notre outil génère dans tous ces cas des EN bien traduites. Ces résultats montrent l'efficacité de notre outil de traduction et ceci grâce aux étapes de post-traitement qui permettent la réorganisation, l'accord et le réajustement des constituants de l'EN et à la priorité donnée à la traduction des noms composés.

# Conclusion

Dans ce chapitre, nous avons commencé tout d'abord par donner un aperçu sur la plateforme linguistique NooJ et sur l'éditeur de commande noojapply. En effet, dans cette plateforme nous avons construit toutes les ressources qui permettent la reconnaissance et la traduction des EN. En outre, grâce à noojapply nous avons pu exploiter le résultat de chaque étape dans l'étape qui suit et garantir le fonctionnement séquentiel et automatique des étapes de l'outil. Ensuite, nous avons présenté la conception de notre outil développé. Cette conception est basée sur le langage UML. Enfin, nous avons expérimenté et évalué notre travail en se basant sur les métriques d'évaluation pour la reconnaissance et nous avons, aussi, effectuer une étude comparative entre notre outil développé et les traducteurs Google et Babylon pour la phase de traduction. L'étude expérimentale et l'évaluation effectuée nous a permis de prouver la faisabilité de notre système et de discerner ses limites.

Dans ce travail, nous avons pu séparer la reconnaissance des EN de leur traduction. Ainsi, si nous voulons par exemple traduire les EN arabes vers une autre langue autre que le français, le module de reconnaissance peut être réutilisé avec quelques modifications si c'est nécessaire (liées aux spécificités du domaine traité). De plus, si nous désirons traduire de n'importe quelle langue vers le français, aussi le module traitant la traduction peut être réutilisé. En effet, ce module traite les spécificités de la langue française.

# Conclusion générale

Dans le présent travail, nous avons développé un système de reconnaissance des EN arabes, en particulier pour les noms de lieux sportifs, et leur traduction vers le français. La phase de reconnaissance est basée sur un modèle de représentation formelle des EN arabes. Ce modèle, qui est proposé à travers une étude typologique effectuée sur un corpus d'étude du domaine du sport, a permis l'identification des dictionnaires et des grammaires nécessaires pour la reconnaissance des noms de lieux sportifs.

La phase de traduction est effectuée en quatre étapes :

- la traduction mot à mot,
- la réorganisation et accord,
- le réajustement et
- la translittération qui s'effectue uniquement pour les mots auxquels aucune traduction n'est affectée. Cette translitération repose sur un ensemble de grammaires morphologiques.

Notons que les phases de reconnaissance et de traduction ont été volontairement séparées ; c'est-à-dire que la traduction n'est entamée qu'une fois la reconnaissance est achevée. L'objectif recherché est l'amélioration du taux de réutilisation des grammaires conçues. En effet, la phase de reconnaissance peut être intégralement réutilisée si la langue cible change.

L'expérimentation des deux phases nous a permis d'une part, de tester la faisabilité du système réalisé et d'autre part, de discerner les limites rencontrées. Les métriques d'évaluation : F-mesure, Précision et Rappel nous ont permis d'évaluer la phase de reconnaissance. L'étude comparative avec d'autres traducteurs reconnus supportant la langue arabe tels que Google et Babylon nous ont montré que le prototype a donné des résultats satisfaisants.

Les principales contributions de ce travail peuvent se résumer dans les points suivants :

- La réalisation d'une étude sur l'état de l'art permettant de retenir une définition appropriée à notre travail de la notion d'EN, de proposer une hiérarchie de type relative au domaine du sport, de choisir les formalismes adéquats à la tâche de reconnaissance et d'identifier les lacunes existantes dans la traduction des EN afin de les palier.

- la proposition d'un modèle typologique permettant le typage des principaux constituants d'une EN donnée à partir des catégories prédéfinies, de prévoir l'effet de l'imbrication des EN sur le processus de leur reconnaissance et de prouver qu'une modélisation à travers une liste de traits peut être une solution appropriée pour représenter explicitement la notion d'emboitement des EN.
- la proposition d'un modèle de représentation formelle des EN, inspiré de certains cadres formels, permettant de vérifier certains principes et contraintes (ex. : saturation et non saturation), de distinguer les différentes ressources (dictionnaires, transducteurs morphologiques et syntaxiques) et de les enrichir.
- la proposition d'une démarche pour la reconnaissance des EN arabes et en particulier les noms des lieux du domaine du sport tout en montrant que les transducteurs à états finis et les réseaux de transitions augmentés (ATN) peuvent jouer un rôle important dans la résolution de plusieurs problèmes linguistiques et fournissent des solutions élégantes permettant d'alléger le processus d'analyse.
- la proposition d'une démarche pour la traduction des EN déjà reconnues vers le français utilisant des heuristiques et des traitements spécifiques garantissant une certaine souplesse à travers l'intégration d'un module de translittération.

Les solutions préconisées tiennent compte de la qualité des ressources linguistiques informatisées en termes d'extensibilité et d'interopérabilité. L'extensibilité et l'interopérabilité sont garanties par l'utilisation de formalismes standardisés. En effet, les ressources (lexicales et syntaxiques) que nous avons conçues sont basées sur les transducteurs à états finis.

Actuellement, nous sommes impliqués dans la création d'une nouvelle norme ISO pour les Entités Nommées. Notre tâche est de contribuer dans définition d'une représentation abstraite et simplifiée de l'EN arabe. La représentation permettrait de normaliser les Entités Nommées tout en donnant des solutions aux problèmes rencontrés et implémentant les différentes contraintes de nature à garantir l'unicité d'une occurrence d'EN. Le but est d'avoir un noyau commun, à toutes les EN, qui offre des extensions pour raffiner la spécification selon la langue et le domaine.

Comme perspectives, nous comptons généraliser le processus de traduction de certains types d'EN (p. ex., noms d'organisations) au niveau de la hiérarchie du domaine et aussi au niveau de son traitement indépendamment du domaine. Nous visons aussi à améliorer notre cadre de représentation par l'introduction d'autres traits sémantiques.

# Bibliographie


ACE Pilot Study Task Definition, 2000. Entity Detection and Tracking - Phase 1. In *ACE*., 2000.

ACE, 2005. The ACE 2005 (ACE05) Evaluation Plan Evaluation of the Detection and Recognition of ACE Ent, K., 2002. Translating Named Entities Using Monolingual and Bilingual Resources. In *P* ities, Values, Temporal Expressions, Relations and Events., 2005.

Al-Onaizan, Y. & Knight*roceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia, 2002.

Bauer, G., 1985. *Namenkunde des Deutschen*.

Beesley, K.R. & Karttunen, L., 2003. Finite State Morphology. *CSLI (Studies in Computational Linguistics)*.

Ben Hamadou, A., Piton, O. & Fehri, H., 2010. Recognition and translation Arabic-French of Named Entities: case of the Sport places. *CoRR abs/1002.0481: 2010*.

Benajiba, Y., 2009. *Arabic Named Entity Recognition*. Thèse de doctorat. Valencia, Spain: Université Politécnica de Valencia.

Bourigault, D., 2002. Upery: un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *TALN'2002*. Nancy, 2002.

Bresnan, J., 2001. Lexical Functional Syntax..

Chinchor, N., 1998. OVERVIEW OF MUC-7/MET-2. In *Proceedings of Seventh Message Understanding Conference (MUC-7)*. Fairfax, Virginia, 1998.

Coates-Stephens, S., 1992. The Analysis and Acquisition of Proper Names for the Understanding of Free Text. *COMPUTERS AND THE HUMANITIES*, 26(5-6), pp.441-56.

Constant, M., 2003. *Grammaires locales pour l'analyse automatique de textes*. Thèse de doctorat. Paris: Universite de Marne-la-Vallée.

Daille, B., Fourour, N. & Morin, E., 2000. Catégorisation des noms propres: une étude en corpus. *Cahiers de grammaire*, pp.115-29.

Ehrmann, M., 2008. *Les entités nommées, de la linguistique au TAL : statut théorique et et méthodes de désambiguïsation*. Thèse de doctorat. Université Paris 7.

Elkateb-Gara, F., 2004. *Extraction d'entités nommées*. Séminaire LIR (Groupe Langue, Information et Représentation).

Enjalbert, P., 2005. L'extraction d'information. *Sémantique et traitement automatique des langues, Hermès Sciences, Lavoisier*, pp.309-34. Traité IC2, série Cognition et traitement de l'information.

Fehri, H., Haddar, K. & Ben Hamadou, A., 2008. Reconnaissance automatique et analyse sémantique d'entités nommées en Arabe. In *actes de laconférence nationale GEI 2008 nouvelles tendances technologies en génie électrique et informatique*. Sousse, Tunisie, 2008. CPU.

Fehri, H., Haddar, K. & Ben Hamadou, A., 2009a. Integration of a transliteration process into an automatic translation system for named entities from Arabic to French. In *proceedings of NooJ'09 conference*. Tozeur, Tunisia, 2009a.

Fehri, H., Haddar, K. & Ben Hamadou, A., 2009b. Translation and Transliteration of Arabic Named Entities. In *proceedings of 4th Language & Technology Conference*. Poznań, Poland, 2009b.

Fehri, H., Haddar, K. & Ben Hamadou, A., 2010a. *Automatic Recognition and semantic Analysis of Arabic Named Entities*. Budapest: Cambridge Scholars Publishing.

Fehri, H., Haddar, K. & Ben Hamadou, A., 2010b. Proposal of a framework for the representation of Arabic named entities to use the transfer approach with NooJ. In *NooJ 2010*. Komotini, Greece, 2010b.

Fehri, H., Haddar, K. & Ben Hamadou, A., 2011a. Separation of recognition of Arabic named entities with NooJ. In *NooJ 2011*. Dubrovnik, 2011a.

Fehri, H., Haddar, K. & Ben Hamadou, A., 2011b. Recognition and Translation of Arabic Named Entities with NooJ Using a New Representation Model. In *FSMNLP 2011*. France, 2011b.

Fehri, H., Haddar, K. & Ben Hamadou, A., 2011c. A new representation model for the automatic recognition and translation of Arabic Named Entities with NooJ. In *RANLP 2011*. Hissar, Bulgaria, 2011c.

Fourour, N., 2002. Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. In *In Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'02)*. Nancy, France, 2002.

Fourour, N. & Morin, E., 2003. Apport du Web dans la reconnaissance des entités nommées. *Revue Québécoise de Linguistique (RQL)*, (32(1)), p.41–60.

Frantzi, K., 1998. *Automatic Recognition of Multi-Word Terms*. Thèse de doctorat. England: Manchester Metropolitan University.

Friburger, N., 2002. *Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques*. Thèse de doctorat. Université François Rabelais Tours.

Friburger, N. & Maurel, D., 2004. Finite-state transducer cascade to extract named entities in texts. *Theoretical Computer Science*, 313 , pp.94 -104.

Gerald, G., Klein, E.H., Pullum, G.K. & Sag, I.A., 1985. Generalized Phrase Structure Grammar.

Gornostay, T. & Skadiņa, I., 2009. Pattern-based English-Latvian Toponym Translation. In *European Association for Machine Translation conference*., 2009.

Grass, T., 2000. Typologie et traductibilité des noms propres de l'allemand vers le français. *Traitement automatique des noms propres, TAL*, 41(3), pp.643-69.

Grishman, R., 1995. The NYU system for MUC-6 or where's syntax? I. In Publishers, M.K., ed. *In Sixth message understanding conference MUC-6*. S. Francisco (CA), 1995.

Grishman, R. & Sundheim, B., 1996. Message Understanding Conference - 6: A Brief History. In *Proceedings of the 16th conference on Computational linguistics (COLING-96)*. Copenhagen, 1996.

Gross, M., 1993. Local grammars and their representation by finite automata. In M. Hoey, ed. *Data, Description, Discourse, Papers on the English Language in honour of John McH Sinclair*. Londres: Harper-Collins. pp.26-38.

Gross, G., 1996. *Les expressions figées en français: Noms composés et autres locutions*. Paris: OPHRYS.

Gross, M., 1997. The construction of local grammars. In E. Roche & Y. Schabes, eds. *Finite-State Language Processing, Language, Speech, and Communication*. MIT Press. Ch. 11. p.329–354.

Gross, M., 1999. Lemmatization of compound tenses in English = Lemmatisation des temps composés en anglais. *Lingvisticae investigationes*, 22, pp.71-122.

Jonasson, K., 1994. *Le Nom Propre. Constructions et interprétations*. Louvain-la-Neuve, Belgique.

Lavecchia, C., 2010. *Les triggers inter-langues pour la Traduction Automatique Statistique*. Thèse de doctorat. Nancy 2.

Le Meur, C., Galliano, S. & Geoffrois, E., 2004. *Conventions d'annotations en Entités Nommées - ESTER* -. [Online].

Ling, W. et al., 2011. Named Entity Translation using Anchor Texts. In *IWSLT*., 2011.

Maurel, D., 1990. Adverbes de date : Etude préliminaire à leur traitement automatique. *Lingvisticae investigationes* , 14(1), pp.31-63.

Maynard, D. & Ananiadou, S., 1999. Identifying contextual information for multi-word term extraction. In *Proc. 5 th Int. Congress on Terminology and Knowledge Engineering (TKE 99)*. Innsbruck, Austria, 1999.

Meilland, J.-C. & Bellot, P., 2005. Extraction automatique de terminologie à partir de libellés textuels courts. *La Linguistique de corpus*, pp.357- 370.

Mesfar, S., 2008. *Analyse morpho-syntaxique automatique et reconnaissance des entités nommées en arabe standard*. Thèse de doctorat. Besançon, France: Université de Franche-Comté.

Morril, G., 1995. Discontinuity in categorial grammar. *Linguistics and Philosophy*, pp.175-219.

MUC-6, 1995. *Named Entity Task Definition. Version 2.0.* [Online].

NLM, 1997. UMLS Knowledge Sources., 1997. U.S. Dept of Health and Human Services, 8è édition.

Paik, W., Liddy, E.D., Yu, E. & McKenna, M., 1996. Categorizing and standardizing proper nouns for efficient information retrieval. *Corpus processing for lexical acquisition*, pp.61 - 73.

Poibeau, T., 2003. Extraction automatique d'information. Du texte brut au web sémantique. *Hermès*, p.250 pages.

Poibeau, T., 2005. Sur le statut référentiel des entités nommées. In *Actes de la conférence Traitement Automatique des Langues Naturelles*. Dourdan. France, 2005.

Pollard, C. & Sag, I.A., 1987. *Information-based syntax and semantics*. Stanford, CA: Center for the Study of Language and Information.

Pollard, C. & Sag, I.A., 1994. *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

Rangel Vicente, M., 2005. La glose comme outil de désambiguïsation référentielle des noms propres purs. *CORELA - Le traitement lexicographique des noms propres*.

Roux, M., EL Zant, M. & Royauté, J., 2006. Projet EPIDEMIA Intervention des transducteurs Nooj. In *IX INteX/Nooj conference 2006. (book of abstracts)*. Belgrade, serbia, 2006.

Sekine, S., 2004. *Named Entity: History and Future*. [Online] Available at: http://nlp.cs.nyu.edu/sekine/Main/publications.html.

Sekine, S. & Eriguchi, Y., 2000. Japanese Named Entity Extraction Evaluation - Analysis of Results. In *Proceedings of the International Conference on Computational Linguistices*. Saarbruecken, Germany, 2000.

Sekine, S. & Isahara, H., 1999. IREX project overview. In *Proceedings of the IREX workshop*. Tokyo Japan, 1999.

Sekine, S. & Nobata, C., 2004. Definition, Dictionary and Tagger for Extended Named Entities. In *Proceedings of the Forth International Conference on Language Resources and Evaluation*. Lisbon, Portugal, 2004.

Sekine, S., Sudo, K. & Nobata, C., 2002. Extended Named Entity Hierarchy. In *The Third International Conference on Language Resources and Evaluation*. Canary Island, Spain, 2002.

Shaalan, K. & Raza, H., 2009. NERA: Named Entity Recognition for Arabic. *Journal of the American Society for Information Science and Technology*, Volume 60 Issue 8, pp.1652-63.

Silberztein, M., 1993. *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*. Informatique linguistique ed. Paris.

Silberztein, M., 2003. Finite-State Description of the French Determiner system. *Journal of French Language Studies*, 13(02), pp.221-46.

Silberztein, M., 2005. NooJ: a Linguistic Annotation System for Corpus Processing. In *HLT/EMNLP*., 2005.

Silberztein, M. & Tutin, A., 2005. NooJ, un outil TAL pour l'enseignement des langues. Application pour l'étude de la morphologie lexicale en FLE. *spécial Atala*, 8(2), pp.123-34.

Silbeztein, M., 2004. NooJ : an Object-Oriented Approach. In Muller, C., Royauté, J. & Silberztein, M., eds. *INTEX pour la Linguistique et le Traitement Automatique des Langues. Proceedings of the 4th and 5th INTEX workshop*. Besançon, 2004. Presses Universitaires de Franche-Comté.

Taghva, K. & Gilbreth, J., 1999. Recognizing acronyms and their definitions. *INTERNATIONAL JOURNAL ON DOCUMENT ANALYSIS AND RECOGNITION*, 1(4), pp.191-98.

Tjong Kim Sang, E.F., 2002. Introduction to the CoNLL-2002 Shared Task : Language-Independent Named Entity Recognition. In Roth, D. & van den Bosch, A., eds. *Proceedings of CoNLL-2002*. Taipei, Taiwan, 2002.

Tjong Kim Sang, E.F. & De Meulder, F., 2003. Introduction to the CoNLL-2003 Shared Task:Language-Independent Named Entity Recognition. In Daelemans, W. & Osborne, M., eds. *Proceedings of CoNLL-2003*. Canada, 2003. Edmonton.

Tran, M., 2006. *Prolexbase. Un dictionnaire relationnel multilingue de noms propres : conception implantation et gestion en ligne*. Thèse de doctorat. Université François Rabelais Tours.

Weissenbacher, D., 2003. *Etude et reconnaissance automatique des relations de synonymie et de renommage dans les textes de génomique*. Master's Thesis. Paris XIII.

Wolinski, F., Vichot, F. & Dillet, B., 1995. Automatic processing of proper names in texts. In *EACL '95 Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*. San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

Woods, W.A., 1970. Transitive network grammars for natural language analysis. *Communications of the ACM*, 13, p.591–606.

# Annexe 1 : Grammaires locales et transducteurs

Les machines à états finis ont été jugées comme les plus adéquates pour les descriptions morphologiques et phonologiques de toutes les langues, indépendamment de leurs natures (romanes, sémitiques, germaniques, etc.) (Beesley & Karttunen, 2003). En effet, il existe un certain nombre d'avantages qui rendent la technologie à états finis particulièrement attractive dans le domaine du traitement automatique des langues naturelles connu par l'acronyme TALN. Parmi ces avantages, nous citons la représentativité, la réversibilité, la modularité, la rapidité et l'efficacité des traitements.

Dans la présente annexe, nous présentons les grammaires formelles tout en précisant leur classification. Ensuite, nous donnons une idée sur le modèle équivalent pour la représentation de chaque type de grammaire. Nous donnons aussi l'intérêt de chaque modèle dans le TALN. Enfin, nous décrivons les spécificités des grammaires locales utilisées par la plateforme linguistique NooJ.

## Les grammaires formelles

Une grammaire est un formalisme qui permet de définir une syntaxe et donc un langage formel. La notion de grammaire formelle est particulièrement utilisée en programmation logique, compilation (analyse syntaxique), en théorie de la calculabilité et dans le traitement des langues naturelles (tout ce qui concerne essentiellement leur morphologie et leur syntaxe).

A l'issue d'une analyse morphosyntaxique, contenant un certain nombre d'informations comme la catégorie grammaticale, le lemme, les flexions et d'autres connaissances il s'avère nécessaire d'étudier comment les mots se combinent pour former des syntagmes puis des propositions et enfin des phrases correctes. Une étude pareille ne peut être effectuée que dans le sein d'une approche syntaxique qui décrit comment ces éléments s'ordonnent pour créer des constituants et composer par la suite des phrases. La syntaxe occupe une place centrale en TAL, en fait, elle constitue une phase presque obligée.

Historiquement, nous noterons que, après l'échec reconnu de la traduction automatique, les algorithmes d'analyse syntaxique sont devenus, pendant les années 60, l'axe des recherches

en traitement automatique. Cela en lien avec l'importance prise par la formalisation et la mathématisation de la syntaxe (d'où le terme des grammaires formelles), importance provenant des travaux de Chomsky. A la fin des années 50, les écrits de N. Chomsky (la publication du livre *« syntactic structures »*1957) ont marqué la théorie des grammaires formelles et ont constitué un point de repère en linguistique informatique.

Par définition, une grammaire formelle est l'ensemble de règles parfaitement explicites, applicables de façon mécanique, qui transforme une entrée en une sortie particulière.

Une grammaire formelle (G) est composée :

- De catégories syntaxiques intitulés éléments non terminaux (Vn) représentées en majuscules (comme (N) pour nom, (SN) pour syntagme nominal).
- D'un ensemble de mots ou bien de vocabulaire terminal (Vt) (ce sont les éléments du lexique) représenté en minuscules.
- D'un ensemble de règles de réécriture, dites aussi règle de production (R) qui spécifient comment une catégorie syntaxique peut être décomposée en une séquence d'autres symboles (catégories ou mot).
- D'un élément distingué, appelé **axiome (P)**, symbolisant la phrase. Une grammaire n'a toujours qu'un seul symbole initial.

L'application successive de règles de productions s'appelle une dérivation. Chaque production doit avoir exactement un symbole non terminal à gauche et à droite ; soit un symbole terminal unique, soit un symbole terminal suivi d'un non terminal. Le langage défini par une grammaire est l'ensemble des mots formés uniquement de symboles terminaux qui peuvent être atteints par une suite de dérivation à partir de l'axiome.

Chomsky engendre une hiérarchie des grammaires formelles qui englobe quatre types :

**Les grammaires de types 0 :** Sont appelés aussi énumérables ou calculables, ces grammaires n'imposent aucune restriction sur les règles de production, qui peuvent être composées à droite et à gauche d'un nombre indéterminé de symboles terminaux et non terminaux. Ce type de grammaire peut résoudre plusieurs problèmes des langues naturelles mais il ne peut être reconnu que par des mécanismes très puissants.

**Les grammaires de type 1 :** sont les grammaires sensibles au contexte ou dépendantes du contexte. Ce type de grammaire impose une restriction sur la taille des règles : le côté droit de la règle ne peut pas avoir moins de symbole que le côté gauche.

**Les grammaires de type 2 :** sont appelées aussi indépendantes du contexte puisque chaque production doit avoir un seul symbole non terminal à gauche et une chaîne non nulle à droite.

**Les grammaires de types 3 :** se sont les grammaires régulières ou bien les grammaires à état finis, ce type de grammaire décrit une opération de concaténation, c'est-à-dire qu'elles désignent que les éléments sont mis en chaîne dans un ordre linéaire pour constituer un syntagme.

Il y a donc 4 grandes familles de grammaires, emboîtées les unes dans les autres. Il est en effet facile de se convaincre que les critères qui caractérisent chacune de ces familles sont de moins en moins restrictifs : toute grammaire qui vérifie celui d'une certaine classe vérifie aussi nécessairement ceux des classes de type inférieur. On peut ainsi visualiser les ensembles de grammaires d'un type donné par la **Figure 57**.

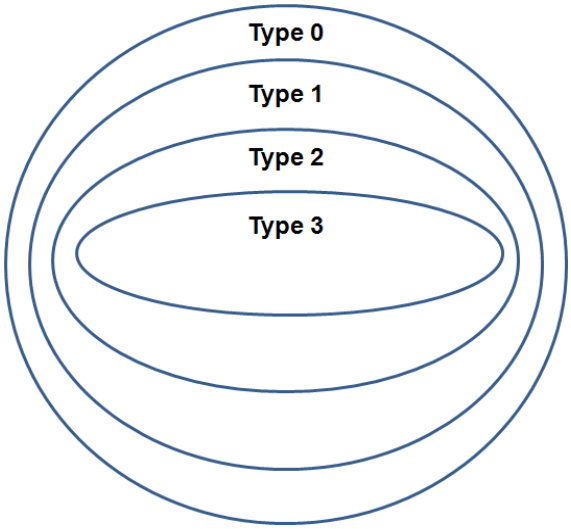


**Figure 57.** *Les classes de grammaires de la hiérarchie de Chomsky*

Outre les quatre types de la hiérarchie de Chomsky, il existe des classes intermédiaires remarquables, par exemple :

- entre 3 et 2 : les langages algébriques déterministes, reconnaissables par automate à pile déterministe ; une classe plus large est la famille des langages algébriques inambigus.
- entre 1 et 0 : les langages récursifs, c'est-à-dire reconnaissables par une machine de Turing (celle-ci doit refuser les mots qui ne sont pas du langage).

Les six types ci-dessus sont strictement inclus les uns dans les autres. Si dans le type 1 on transforme « non déterministe » en « déterministe » on obtient un type plus petit mais on ne sait pas montrer s'il est strictement inclus dans le type 1 ou s'il est égal à celui-ci.

Nous avons vu que chaque grammaire impose une restriction sur les productions autorisées. Elles caractérisent des langages différents et ne présentent pas la même génération.

D'après cette variété et cette multiplicité des grammaires une question s'impose est de savoir quel type de ces grammaires peut effectuer un traitement satisfaisant de la langue naturelle.

Les grammaires les plus utilisées pour le traitement de la langue naturelle sont les grammaires de type 2 indépendantes du contexte. En effet, la grande majorité de phénomènes langagiers peut être décrite à l'aide de grammaires indépendantes du contexte qui constituent un moyen rigoureux de fonder la syntaxe des langages formels et elles présentent un certain nombre de propriétés intéressantes pour le traitement de la langue naturelle. Il est vrai que les grammaires indépendantes du contexte sont les plus utilisées, puisqu'elles permettent de décrire la plupart des structures syntaxiques de la langue naturelle et aussi de reconnaître les phrases correctes. Néanmoins, ce type de grammaire est insuffisant pour résoudre plusieurs problèmes comme « l'accord », les dépendances non bornées (les éléments qui sont éloignés les uns des autres) comme la négation et l'interrogation. Face à ces insuffisances, l'appel et le besoin à d'autres grammaires sont devenus une nécessité urgente, d'où la naissance d'autres grammaires et d'autres théories qui veulent combler les carences et les insuffisances de la grammaire Chomskyenne. Nous trouvons essentiellement les grammaires d'unification qui se sont développées en réaction au courant transformationnel de Chomsky.

# Expressions régulières

Dans le traitement des textes écrits, il est souvent nécessaire de rechercher dans ces textes des segments ayant une caractéristique particulière : les phrases, les mots, les entités nommées, un mot particulier, les formes fléchies d'un lemme particulier etc. L'utilisation des expressions régulières constitue un moyen d'effectuer une telle tâche. En effet, une expression régulière est une formule permettant de caractériser un ensemble de chaînes de caractères.

## Définition

Formellement, une expression régulière (ou rationnelle) sur un alphabet de symboles $\Sigma$ est une expression construite à partir de symboles de $\Sigma$ et du mot vide $\varepsilon$, à l'aide de trois opérations : la concaténation (les expressions concaténées sont simplement juxtaposées), la

disjonction « | » et la clôture de Kleene « * ». Elle définit un ensemble de chaînes de symboles de Σ qui forme un langage régulier.

A partir des trois opérations primitives, on définit habituellement deux opérations complexes qu'on exprime à l'aide des opérateurs ? et + :

*Si E est une expression régulière,* $E? = E \mid \varepsilon$ *et* $E+ = E\,E*$

Dans certains langages formels comme Python, on utilise d'autres opérations complexes :

| | |
|---|---|
| **.** | Remplace n'importe quel symbole (sauf éventuellement le caractère spécial \n de passage à la ligne) |
| **[ ]** | Remplace l'un quelconque des symboles placés entre les crochets |
| **[^ ]** | Remplace l'un quelconque des symboles qui ne sont pas entre les crochets |
| **{m,n}** | m et n sont des entiers qui indiquent respectivement le nombre minimum et maximum de fois que la sous-expression située juste avant est répétée |
| **\w** | Remplace n'importe quel caractère alphanumérique (lettre ou chiffre) plus le caractère _ |
| **\W** | Remplace n'importe quel symbole qui n'est ni un caractère alphanumérique, ni le caractère _ |
| **\d** | Remplace n'importe quel chiffre |
| **\D** | Remplace n'importe quel symbole qui n'est pas un chiffre |
| **\s** | Remplace l'un quelconque des caractères d'espacement ' ', \t, \n, \f, \r \v |
| **\S** | Remplace n'importe quel symbole qui n'est pas un caractère d'espacement |

Dans des expressions régulières complexes, les opérations unaires qui sont toutes postfixées, sont effectuées en premier. Ensuite, ce sont les concaténations et enfin les disjonctions. Les parenthèses permettent de forcer les priorités.

# Utilisation

L'application la plus répandue des expressions régulières en TAL consiste à rechercher dans un texte des chaînes de caractères ayant une caractéristique donnée. Cette caractéristique est exprimée par une expression régulière.

La plupart des algorithmes utilisés parcourt le texte du début à la fin en essayant de trouver la plus longue chaîne de caractères correspondant à l'expression régulière. Dès qu'une chaîne est trouvée, le processus est réitéré, à partir du caractère, suivant la fin de la chaîne trouvée.

Pour exprimer certaines contraintes sur la position des chaînes de caractères recherchées, la syntaxe des expressions régulières est étendue par des constantes de contrôle ou ancres. Les ancres ne représentent aucune chaîne de caractères (le mot vide) mais elles indiquent des positions particulières dans le texte.

| ^ | **Début d'une ligne** |
|---|---|
| **$** | Fin d'une ligne |
| **\b** | Début ou fin d'un mot |
| **\B** | Position qui n'est ni un début, ni une fin de mot |
| **\A** | Début du texte |
| **\Z** | Fin du texte |

## Applications au TAL

Les expressions régulières jouent un rôle essentiel pour segmenter un texte en phrases et en mots. Elles permettent de détecter grossièrement les séparateurs de phrases et de mots. En outre, elles servent à rechercher des mots dans les éditeurs de textes.

En recherche d'informations, elles permettent de retrouver des documents et des informations à l'intérieur de ces documents. Elles sont aussi utilisées pour indexer des documents. Elles peuvent être aussi employées dans l'analyse morphologique et l'étiquetage morpho-syntaxique.

# Les grammaires locales

La notion de grammaire locale est développée à partir des années 80 par Maurice Gross. Elles servent à localiser des phénomènes locaux de manière très précise dans les textes, comme les dates (D. Maurel, 1990), les déterminants numéraux (M. Silberztein 1993, A. Chrobot, 2000), les incises (C. Fairon, 2000). Elles sont équivalentes à des réseaux récursifs de transitions [RTN] (W. Woods, 1970). Elles ont le grand avantage de pouvoir être appliquées directement et efficacement à des textes (M. Silberztein, 1993). En effet, les grammaires locales,

constituées d'automates finis et couplées aux dictionnaires morpho-syntaxiques, permettent l'analyse automatique de textes par le logiciel Intex développé par Max Silberztein.

# Les automates finis

Les limites d'une expression régulière est de détecter dans un texte les chaînes de caractères qui appartiennent au langage défini par cette expression. Pour résoudre d'une façon automatique ce problème, il est nécessaire d'avoir un mécanisme de reconnaissance de ce langage. Dans ce contexte, on trouve les automates d'états finis qui sont des machines abstraites destinées à reconnaître un langage régulier, c'est-à-dire un langage défini par une expression régulière. Il peut être normalisé pour avoir une grande efficacité en temps et en espace de calcul.

## Définition

Les automates finis sont des « machines abstraites » qui savent reconnaître l'appartenance ou la non-appartenance d'un mot à un langage régulier donné. Ces machines abstraites constituent un modèle théorique de référence.

Les applications qui implémentent la notion d'automates finis ou ses variantes sont nombreuses. Un automate « lit » un mot écrit sur son ruban d'entrée. Il part d'un état initial et à chaque lettre lue, il change d'état. Si, à la fin du mot, il est dans un état final, on dit qu'il reconnaît le mot lu.

Les automates peuvent être combinés de manières très variées pour construire des automates plus complexes.

Un automate d'états finis est composé de deux parties :

- Un ruban infini avec des positions numérotées 0, 1, … pouvant contenir chacune un symbole d'un alphabet donné. Le ruban est muni d'un pointeur sur la position courante qui peut être seulement lue.
- Une unité de contrôle qui pilote l'avancée pas à pas du pointeur sur le ruban en fonction de son état et du symbole lu sur le ruban.
- Un automate d'états finis sur un alphabet $\sum$, est un 5-uplet $(Q, \sum, q_0, F, \delta)$ tel que:
- Q est un ensemble fini d'états.

- $\sum$ est un alphabet fini de symboles. En lui ajoutant le mot vide ε, on obtient l'ensemble des étiquettes d'entrée de l'automate.
- $q_0$ est un élément particulier de Q, l'état initial de l'automate.
- F est une partie de Q rassemblant les états acceptants de l'automate.
- δ est une relation de transition qui associe un état de départ $q_1$ et une étiquette d'entrée $a$ avec un état d'arrivée $q_2$. Ceci est noté : δ ($q_1$, $a$, $q_2$).

Un calcul sur un automate est une suite (éventuellement vide) de transitions de la forme : $q_0 \xrightarrow{a_1} q_1 \xrightarrow{a_2} \ldots \xrightarrow{a_n} q_n$ où $q_0$ est l'état initial.

On note alors : $q_0 \xrightarrow{a_1 a_2 \cdots a_n} * q_n$

Le mot $a_1 a_2 \ldots a_n$ est reconnu par l'automate si $q_n$ est un état acceptant.

Le langage reconnu par l'automate est l'ensemble de tous les mots reconnus par l'automate.

La **Figure 58** est un exemple d'un automate.

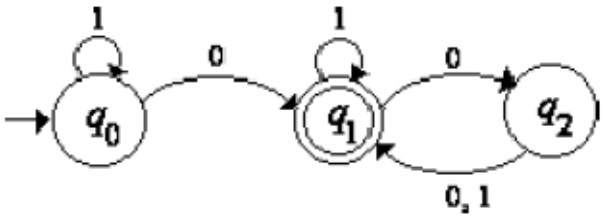


**Figure 58.** *Exemple d'un automate*

Dans l'automate de la **Figure 58**, Q = {$q_0$, $q_1$, $q_2$}, Σ = {0, 1}, F = {$q_1$} et $q_0$ est l'´etat initial. Quant a la fonction de transition, elle est la suivante : δ ($q_0$, 0) = q1, δ ($q_0$, 1) = $q_0$, δ ($q_1$, 0) = $q_2$, δ ($q_1$, 1) = $q_1$, δ ($q_2$, 0) = q1 et δ ($q_2$, 1) = $q_1$

A partir des opérations qui peuvent être appliquées sur les automates à états finis (l'union, la concaténation et la clôture de Kleene), on peut construire un automate d'états finis pour n'importe quelle expression régulière, qui reconnaisse le langage associé. Réciproquement, pour n'importe quel automate d'états finis, on peut trouver une expression régulière dont le langage associé est celui reconnu par l'automate.

## Applications au TAL

Etant donné une forme d'implémentation des expressions régulières, les automates d'états finis peuvent se retrouver dans toutes les applications qui font intervenir des expressions régulières. Ils permettent de représenter les lexiques et les dictionnaires électroniques de

manière compacte avec un accès efficace. Ils ont aussi une manière simple et efficace d'implémenter des formes différentes de traitement des langues : reconnaissance et synthèse de la parole, analyse morphologique, analyse syntaxique …

# Les transducteurs à états finis

Les automates d'états finis sont des machines destinées à reconnaître certains langages. Parfois, il est nécessaire de réaliser une transformation qui traduise un mot d'un langage dans un autre. C'est le rôle des transducteurs d'états finis. Ainsi, un transducteur d'états finis définit une relation entre un langage d'entrée et un langage de sortie.

## Définition

Un transducteur d'états finis sur un alphabet d'entrée $\Sigma_i$ et un alphabet de sortie $\Sigma_o$ est un 6-uplet $(Q, \Sigma_i, \Sigma_o, q_0, F, \delta)$ tel que:

- Q est un ensemble d'états finis.
- $\Sigma_i$ est un alphabet fini de symboles d'entrée.
- $\Sigma_o$ est un alphabet fini de symboles de sortie.
- $q_0$ est un élément particulier de Q représentant l'état initial du transducteur.
- F est une partie de Q représentant les états acceptant du transducteur.
- $\delta$ est une relation de transition qui associe un état de départ $q_1$ de Q et un mot d'entrée $w_i$ de $\Sigma_i^*$ avec un état d'arrivée $q_2$ de Q et un mot de sortie $w_o$ *de* $\Sigma_o^*$. On écrit alors : $\delta(q_1, w_i, q_2, w_o)$.

Un calcul sur un transducteur est une suite éventuellement vide de transitions de la forme : $q_0 \xrightarrow{a_1 : b_1} q_1 \xrightarrow{a_2 : b_2} \ldots \xrightarrow{a_n : b_n} q_n$ où $q_0$ est l'état initial. On écrit alors : $q_0 \xrightarrow{a_1 a_2 \cdots a_n : b_1 b_2 \cdots b_n} * q_n$

Le mot $a_1\, a_2 \ldots\, a_n$ est dit traduit par le transducteur dans le mot $b_1\, b_2 \ldots\, b_n$ si $q_n$ est un état acceptant.

La **Figure 59** est un exemple d'un transducteur à états finis.

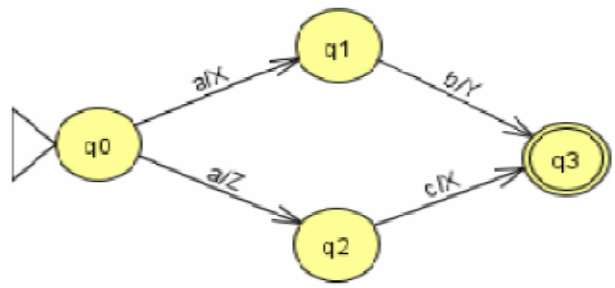


**Figure 59.** *Exemple d'un transducteur*

Le transducteur représenté dans la **Figure 59** permet de reconnaître les deux chaînes de caractères : "ab" et "ac". Il produit, respectivement, les deux sorties : "XY" et "ZX".

La relation réalisée par le transducteur $T$ est l'ensemble des couples $(w_1, w_2)$ tels que $w_1$ est transformé en $w_2$ par le transducteur. Cette relation est appelée une relation régulière (rationnelle).

On associe à tout transducteur $T$ la fonction $|T|$ qui associe chaque mot d'entrée $w_i$ à l'ensemble $|T|$ $(w_i)$ de tous les mots résultant de la transduction de $w_i$ par $T$. Si tout mot d'entrée est transformé en un mot de sortie au plus, le transducteur réalise une fonction rationnelle.

## Applications au TAL

Pour les langues qui n'ont pas une morphologie trop riche (anglais, français…), on peut stocker tous les mots fléchis de la langue dans un lexique morphologique. Une manière efficace de le faire est d'utiliser un transducteur : l'entrée du transducteur reçoit les lemmes et les paramètres de flexion et la sortie des mots fléchis correspondants. L'intérêt est la compacité du lexique, sa rapidité d'accès et la possibilité de l'utiliser tant en génération qu'en analyse.

Pour les langues riches morphologiquement (turc, arabe …), ce sont les règles morphologiques qui sont représentées par des transducteurs.

La représentation d'un lexique morphologique par un transducteur n'est pas en général suffisante pour l'analyse morphologique car elle ne prend pas en compte le fait que les mots fléchis insérés dans des énoncés subissent des modifications causées par les règles phonologiques. On représente quand il est possible chaque règle phonologique par un transducteur. De ce fait, les transducteurs sont composés en cascade en respectant l'ordre d'application des règles.

Les transducteurs peuvent être aussi utilisés pour la segmentation de textes (complétée éventuellement par des règles phonologiques). Ils sont particulièrement efficaces dans l'étiquetage morpho-syntaxique fondé sur des règles de transformation apprises d'un corpus (étiquetage à la Brill).

## Les RTN

Un réseau de transitions récursif (Recursive Transition Network : RTN) est défini par ensemble de graphes semblables à ceux d'un automate fini dans lequel chaque transition permet d'atteindre un état terminal ou non-terminal. La différence par rapport à un automate à états finis se situe au niveau du traitement des états non-terminaux : le RTN traite chaque état non-terminal comme un éventuel appel à d'autres réseaux (y compris le RTN lui-même).

Un RTN peut être défini formellement par un 6-uplet M = (Q, *I*, Σ, δ, $q_0$, F), où :

- Q est un ensemble fini non vide d'états ;
- I est l'ensemble des états sous-initiaux (états qui étiquettent au moins une transition du transducteur RTN, et représentent donc un appel récursif au sous-RTN) ;
- Σ est un alphabet de symboles complexes. Chaque symbole est constitué d'une paire (e,s) avec e ∈ à un alphabet d'entrée E, et s ∈ à un alphabet de sortie S. Σ ⊆ E x S ;
- δ : Q x (Σ ∪ E ∪ {ε}) → Q est la fonction de transition ;
- q0 ∈ Q est l'état initial ;
- F ⊆ Q est l'ensemble non vide des états finaux.

La **Figure 60** est un exemple d'un RTN.

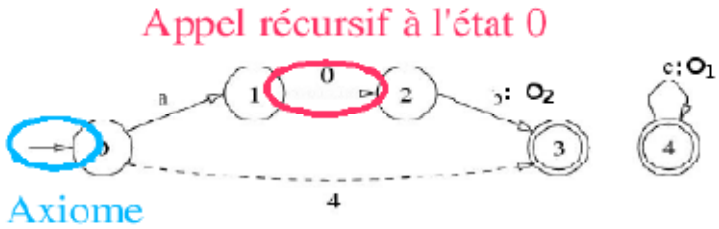


**Figure 60.** *Exemple d'un RTN*

Le transducteur RTN de la **Figure 60** est défini par Q = {0, 1, 2, 3, 4}, I = {0, 4}, Σ = {(a, ε), (b,$O_2$), (c,$O_1$)}, I = {0, 4}, F = {3, 4} et S = {0}. Quelques exemples de séquences reconnues et transformées : ab → $O_2$ accb → $O_1O_1O_2$ c → $O_1$ ...

Un réseau de transitions récursif (RTN) peut être étendu pour donner un réseau de transitions augmenté (ATN : Augmented Transition Network). Un ATN est un RTN auquel s'ajoutent certaines extensions qui lui donnent un pouvoir descriptif supérieur à celui d'une grammaire noncontextuelle.

Trois extensions sont apportées, il s'agit notamment :

- d'ajouter des registres aux réseaux de transitions ;
- d'imposer des conditions sur les transitions ;
- d'associer des actions aux transitions effectuées.

# Les grammaires locales dans NooJ

La plateforme linguistique NooJ fait usage des différentes grammaires locales citées ci-dessus afin de représenter des données, formaliser des phénomènes linguistiques et analyser des textes :

- Les expressions rationnelles représentent un moyen rapide pour les requêtes simples. Par exemple, lorsque la séquence recherchée consiste en quelques mots, il est possible d'énumérer ces mots directement dans une expression rationnelle.
- Les transducteurs à états finis peuvent servir à décrire divers phénomènes linguistiques; notamment pour associer chaque patron retenu à un résultat d'analyse. Ils sont aussi utilisés pour stocker des données lexicalisées ainsi que toutes les informations morpho-syntaxiques qui s'y rattachent.
- Les automates à états finis ne sont qu'un cas particulier des transducteurs : ils produisent le mot vide. Ils servent à localiser des phénomènes morpho-syntaxiques dans un corpus, extraire les séquences reconnues, construire des tables de concordances, etc.
- Les RTN se présentent sous la forme d'ensembles organisés de graphes. Ces graphes peuvent eux-mêmes être des automates, des transducteurs à états finis ou des RTN. Dans la pratique, les RTN sont utilisés pour construire des bibliothèques de graphes facilement réutilisables.
- Les ATN sont des RTN qui contiennent des variables et produisent des contraintes ou des productions complexes. Les variables permettent de stocker les séquences reconnues ; leurs contenus sont ensuite utilisés pour effectuer des transformations.

Bien que théoriquement équivalentes aux définitions formelles ci-dessus citées, NooJ en utilise d'autres pour décrire les différentes machines à états qu'il met en œuvre. En l'occurrence, un transducteur à états finis est défini comme étant un quintuplet T = (N, Σ, C, N0, NT), où:

- N est un ensemble fini non vide de nœuds ;
- Σ est un alphabet de symboles complexes. Chaque symbole est constitué d'une paire (e, s), où
- e ∈ à un alphabet d'entrée E, et s ∈ à un alphabet de sortie S. Σ ⊆ E x S ;
- C est un ensemble fini non-vide de connexions entre les nœuds de N ;
- N0 ∈ N est le nœud initial ;
- NT ∈ N est le nœud terminal;

NooJ propose aussi des façons équivalentes pour construire des machines à états finis. Nous pouvons citer :

- expressions rationnelles équivalentes aux automates ;
- expressions rationnelles à production équivalentes aux transducteurs ;
- grammaires de réécriture équivalentes aux RTN
- grammaires avec variables équivalentes aux ATN.

# Annexe 2 : Interfaces réalisées

L'outil réalisé contient une interface principale interactive, intuitive et conviviale développée dans l'environnement Microsoft Visual C#. En effet, l'exécution du programme provoquera l'apparition de cette interface qui est décrite dans la **Figure 61**.

Dans ce qui suit, nous allons détailler le fonctionnement de notre système via la description des interfaces établies.


**Figure 61.** *Interface d'accueil*

L'interface principale, décrite dans la **Figure 61**, contient une barre de menus qui permet l'accès à toutes les fonctionnalités de notre système. Dans ce qui suit, nous allons détailler le rôle des menus importants.

Le menu *Recognition* permet la reconnaissance des EN. Ainsi, l'activation de ce menu provoquera l'apparition de l'interface de reconnaissance. L'utilisateur doit alors choisir le texte ou corpus désiré en précisant son chemin d'accès et les différentes ressources nécessaires. La validation de son choix, entraine l'affichage des EN reconnues.

Le menu *Translation* permet la traduction des EN. Le choix de ce menu entrainera l'affichage de l'interface de la **Figure 62**.

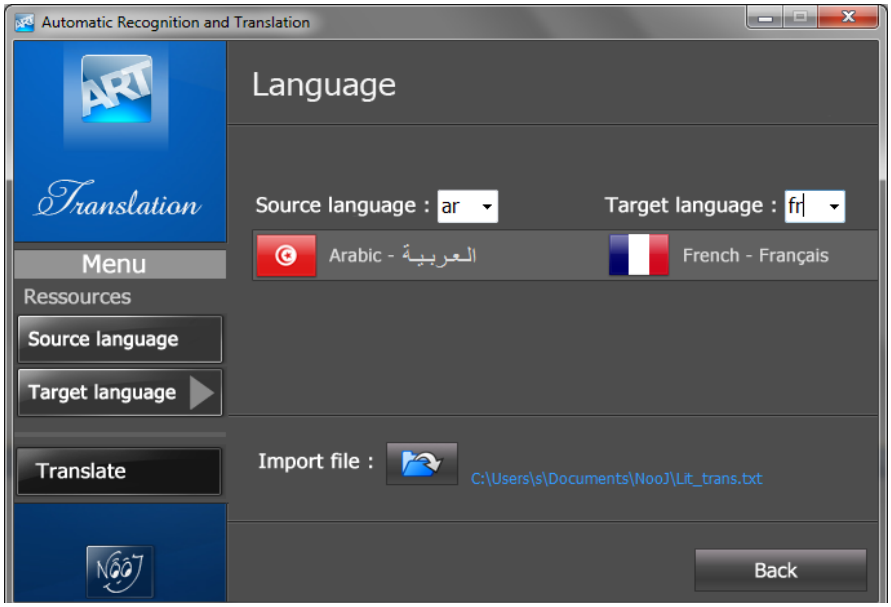


**Figure 62.** *Interface de traduction*

Notons que dans l'interface de la **Figure 62**, l'utilisateur doit choisir la langue source et la langue cible. Dans notre cas, la langue source est l'arabe (ar) et la langue cible est le français (fr). La liste des choix des langues est chargée automatiquement à partir de la plateforme NooJ. L'utilisateur doit aussi préciser le fichier contenant les EN à traduire en précisant son chemin d'accès. Une fois, ces informations sont identifiées l'utilisateur doit choisir les ressources de la langue source c'est-à-dire les ressources de la traduction mot à mot. Ainsi l'interface de la **Figure 63** est affichée.

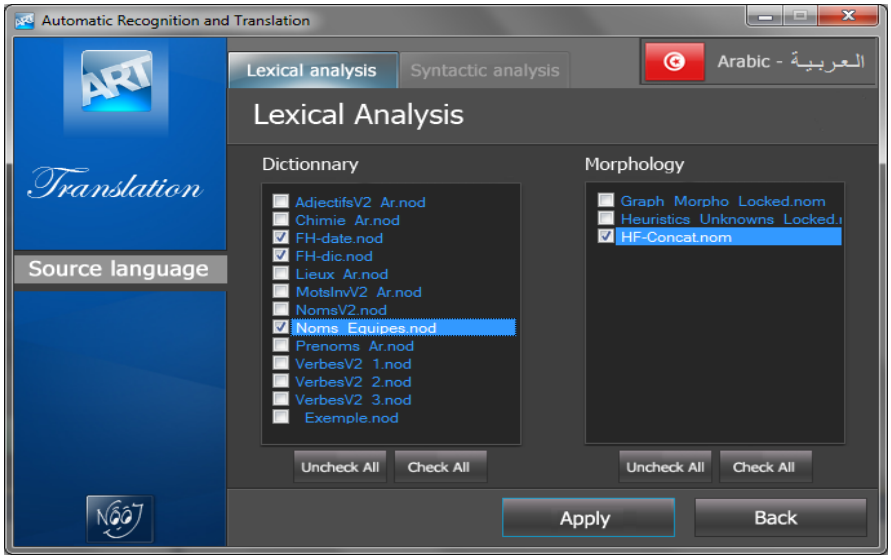


**Figure 63.** *Interface de ressources de la traduction mot à mot*

Les ressources que l'utilisateur doit choisir sont chargées automatiquement selon la langue choisie (ar) à partir de la plateforme NooJ. Ces ressources sont lexicales et syntaxiques. Pour les ressources lexicales, l'utilisateur identifie les dictionnaires et les grammaires morphologiques. Quant aux ressources syntaxiques, l'utilisateur expert identifie les grammaires syntaxiques.

Après avoir sélectionné les ressources de la langue cible, l'utilisateur doit choisir les ressources de la langue cible. Ceci en cliquant sur le bouton *Apply* qui permet de valider le choix effectué et passer à l'étape suivante. Ainsi l'interface de la **Figure 64** est affichée.

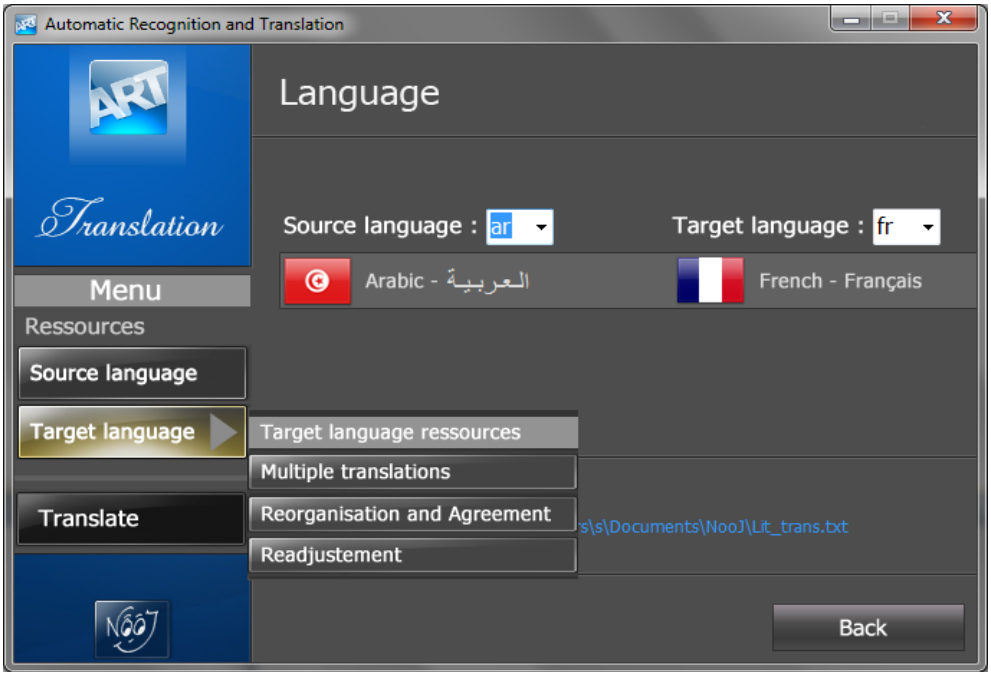


**Figure 64.** *Interface d'identification des ressources de la langue cible*

Remarquons que les ressources de la langue cible (fr) sont celles qui permettent l'élimination des traductions multiples, la réorganisation et accord et le réajustement. L'utilisateur doit alors sélectionner les ressources pour chaque étape. L'interface de la **Figure 65** permet la sélection de celles de l'étape de réorganisation et accord.

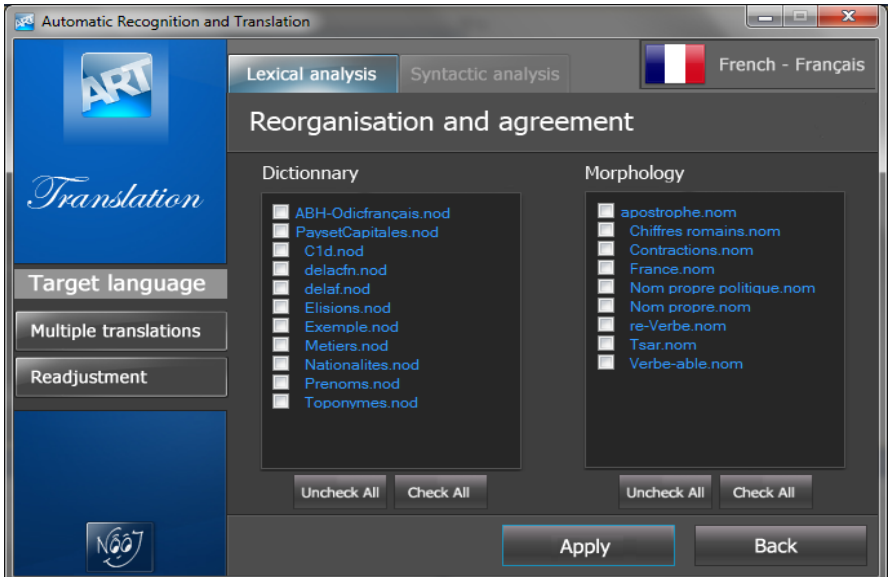


**Figure 65.** *Interface d'identification des ressources de réorganisation et accord*

De la même façon que pour l'identification de la langue source, l'utilisateur doit sélectionner les ressources propres à l'étape de réorganisation et accord. La même procédure s'effectue pour le reste des étapes : traduction multiples et réajustement. Une fois, l'utilisateur valide ses choix, il clique sur le bouton *Translate* de la **Figure 64**. Le retour à cette interface se fait avec le bouton *Apply*. Ainsi, le fichier résultat contenant les EN traduites dans la langue française sera crée et affiché. Notons que l'appui sur le bouton *ART* permet le retour à l'interface d'accueil.

# Annexe 3 : Système de translittération Al-Qalam

## Lettres

| | | | | | |
|---|---|---|---|---|---|
| hamza | ' | | | | |
| 'alef | aa | zayn | z | qaaf | q |
| baa' | b | syn | s | kaaf | k |
| taa' | t | shyn | sh | laam | l |
| thaa' | th | Saad | S | mym | m |
| jym | j | Daad | D | nuwn | n |
| Haa' | H | Taa' | T | haa' | h |
| khaa' | kh | Zaa' | Z | waaw | w |
| daal | d | `ayn | ` | yaa' | y |
| dhaal | dh | ghayn | gh | | |
| raa' | r | faa' | f | | |

| | |
|---|---|
| taa' marbuwTah | t or h |
| haa' marbuwTah | h |
| 'alef maqSuwrah | ae |
| hamzat alwaSl | e |

## Voyellation

| | |
|---|---|
| fatHah | a |
| kasrah | i |
| Dammah | u |
| shaddah | doubler la lettre précédente |
| maddah | ~aa |
| sukuwn | - |
| tanwyn | N |